DOCUMENT RESUME

ED 233 085 TM 830 581

AUTHOR Shepard, Lorrie; And Others
TITLE Accounting for Statistical Artifacts in Item Bias Research.
SPONS AGENCY Colorado Univ., Boulder. Council on Research and Creative Work.; Colorado Univ., Boulder. School of Education.
PUB DATE Apr 83
NOTE 59p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Achievement Tests; Black Students; High Schools; High School Seniors; *Latent Trait Theory; *Racial Bias; Research Methodology; Statistical Analysis; *Statistical Bias; *Test Bias; Testing Problems; Test Items; *Test Validity; White Students
IDENTIFIERS *High School and Beyond (NCES)

ABSTRACT

Theoretically preferred item response theory (IRT) bias detection procedures were applied to both a mathematics achievement and vocabulary test. The data were from black seniors and white seniors on the High School and Beyond data files. We wished to account for statistical artifacts by conducting cross-validation or replication studies. Therefore, each analysis was repeated on randomly equivalent samples of blacks and whites (n's=1500). Furthermore, to establish a baseline for judging bias indices that might be attributable only to sampling fluctuations, bias analyses were conducted comparing randomly selected groups of whites. Also, to assess the effect of mean group differences on the appearance of bias, pseudo-ethnic groups were created, i.e., samples of whites were selected to simulate the average black-white difference. The validity and sensitivity of the IRT bias indices were supported by several findings. The pattern of between study correlations showed high consistency for parallel ethnic analyses where bias was plausibly present. Also, the indices met the discriminant validity test. Overall the sums-of-squares statistics (weighted by the inverse of the variance errors) were judged to be the best indices for quantifying item characteristic curve differences between groups. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED233085

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
L. Shepard
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Accounting for Statistical Artifacts in Item Bias Research[1]

Lorrie Shepard
University of Colorado

Gregory Camilli
Human Systems Institute

David M. Williams
University of Colorado[2]

The question of bias in tests of achievement and mental ability is one of the most prominent issues in both the technical psychometric literature and the more general public concerns about education. Tests are used to make many significant decisions about individuals, for example: to measure school achievements, to assign children to special classes, and to select applicants for prestigious schools and professions. Tests are intended to be more reliable and impartial than subjective judgments about individuals; but if a test is invalid, the decisions derived from it will be unfair. Bias is a kind of invalidity that is disadvantageous to members of certain subgroups that take the test. If a test is biased, critical educational and life decisions may be made unfairly for members of those subgroups.

---

[1]Paper presented at the annual meeting of the American Educational Research Association, Montreal, April 1983.

[2]We wish to thank the Council on Research and Creative Work and Dean Richard Turner, School of Education, University of Colorado, for their financial support.

TM 830 581

Scholars have taken three general approaches to research on the question of test bias (see Cole, 1981; Bond, 1981): (1) predictive validity studies in selection situations; (2) investigation of external biasing factors such as the race of examiner, test-wiseness of examinees, and speed of administration; and (3) construct or content validity studies of the internal structure of the test. The present research is focused on test item-bias methods, which are subsumed in the last category of inquiry. Item-bias methods are statistical procedures intended to test whether items function equivalently in two groups. Therefore, they address the basic validity question: does the test (or individual items in the test) measure what it purports to measure for both groups?

Numerous item-bias methods exist (see Berk, 1982; Rudner, 1980; Shepard, 1981). Most rely on an item-by-group-interaction criterion of bias; that is, statistical adjustments are made for overall group differences, and then items that are relatively more difficult for one group are flagged as potentially biased. For example, an early method proposed by Angoff (1972) was based on transformed item difficulties. The proportion getting each item correct was computed separately for two groups, e.g., blacks and whites. Then, (after a transformation to linearize the relationship) the scattergram depicting the relationship between the two sets of p-values was examined. Most of the data points would fall along a line from the lower left to the upper right-hand corner of the graph, indicating test items that ranged from very easy in both groups to very difficult. Items which deviated substantially from this principal axis line were those that were relatively more difficult in one group than the other, and hence possibly biased.

A standard operating assumption should be discussed regarding item-bias techniques. Because they lack an external criterion, they can only be used to detect relative bias, not pervasive bias in a test (Petersen, 1977). As in the above example, the various methods either use total test score (or estimated abilities from the total set of items), or average p-value differences to define the "typical" difference between groups; this then becomes the standard of "unbiasedness" against which individual items are compared. Thus, if there is bias in the determination of this typical group difference, it will go undetected by these techniques. Despite this limitation, it has been argued that item-bias procedures may be the preferred approach for understanding the nature of bias and for uncovering irrelevant difficulties in items which charge their meaning for members of different groups (Shepard, 1982). The predictive validity models of test fairness involve an external criterion but are not without fault. Petersen and Novick (1976) demonstrated that the various models for defining equal regressions (i.e., equal predictive validity for two groups) are mutually contradictory. Moreover, Linn (1982) has recently explained how differential measurement error for two groups could obscure predictive bias. Finally, there is the actuarial problem (Shepard, 1982). Predictive validity studies look only at the magnitude of the correlation between test and criterion; they do not distinguish between relevant and irrelevant sources of relationship. Nor do they examine whether the combination of predictors that maximize the correlation are equally defensible. Test-item bias methods let us look more directly at what we are measuring. They leave for a second step the question of how measures of separate traits should be combined to make selection or other test-based

decisions.

Among item-bias techniques, the theoretically preferred method is the three-parameter item response theory (IRT) method, also called the three parameter latent-trait or item-characteristic-curve (ICC) method. It is preferred because of its sample-invariant properties that make it less likely that true group differences will be mistaken for bias. Hunter (1975) and Lord (1977) have demonstrated heuristically that bias techniques based on classical test theory (such as p-value differences or point-biserials) will produce invalid indices of bias in the presence of group mean differences. Because p differences interact with item discrimination, items that are merely more discriminating (i.e., better measures of the trait in both groups) will have bigger differences in performance. Furthermore, the variability of a particular group and how "centered" an item is for that group, will artifactually control the item's discriminating power. Methods such as empirical ICCs (Green & Draper, 1972) and chi-square procedures (Scheuneman, 1979) were intended to be approximations to the IRT method. These procedures are crude, however, and will still confound real group differences with bias because of regression effects. The one-parameter latent-trait method (or Rasch model) shares the theoretically sample invariant properties of the three-parameter model. However, the Rasch model is not recommended for bias detection because it will confound other sources of model misfit (particularly differences in item discrimination) with item "bias" (see Divgi, 1981; Ironson, 1982; Shepard, Camilli & Averill, 1981).

The conceptual definition of bias using the three-parameter IRT model and specific procedures will be explained in the method section. The

three-parameter IRT method is applied in this paper because it is theoretically the most sound. Its superiority is relative, however, rather than absolute; bias detection using IRT is not without problems. First there are estimation problems due to sampling fluctuations. Even with reasonably large sample sizes, as in this study, it is possible that misestimated item parameters for separate groups could create or obscure item-characteristic curve differences when the two groups are compared. More importantly there may be larger sources of error when samples are very different. Even the theoretical claims for the model are said to be true only when the model holds. The following discussion is focused on the potential for obtaining invalid bias indices, ever with IRT methods. First, however, a digression is in order regarding substantive interpretation of bias indices. Difficulties encountered when trying to make substantive interpretations of bias analyses may be linked to the problem of statistical artifacts.

Substantive Interpretations of Bias

Given increasing concern over cultural bias in tests, a strong impetus to the development of statistical screening techniques was the apparent failure of judgmental methods for identifying biased test questions. That is, even minority experts, sensitive to the issue of cultural loading in test questions, could not predict with better than chance success what type of items would be difficult for members of particular groups (Jensen, 1977; Plake, 1980; Sandoval & Miille, 1980). For example, Jensen (1976) found that an item on the WISC often cited for its dependence on white middle-class values, "What is the thing to do if a fellow (girl) much smaller than yourself starts to fight with you?", was actually relatively

easier for blacks. If biased test questions were not obvious to expert judges, then perhaps statistical detection procedures could uncover more subtle changes in the meaning of items for different groups.

A more disappointing result--after numerous statistical bias studies--has been that here too expert judges are often at a loss to explain the source of bias in items with large bias indices. For instance, in an early study Lord (1977) found that 46 of 85 items on the verbal SAT were significantly different for black and whites (bias was sometimes against whites). But, in studying the items identified as biased no particular insights could be gained to explain the differential performance. It had been hoped that the use of statistical bias techniques would lead to substantive generalizations about the nature of items found to be biased against specific groups. For example, Scheuneman (1979) found that negatively worded items were biased against blacks. This type of consistent finding turned out to be more the exception than the rule. Raju (in Green et al., 1982) described the serious problems faced by test publishers who may decide to discard statistically deviant items even though they are unable to explain why they are biased "in terms of the content." Scheuneman (1982) best summarized researchers' disappointment with initial efforts to interpret bias:

> We naively assumed that a review of such items would readily reveal the source of apparent bias, that the problem could then be easily corrected with suitable modifications or by dropping the item from the test or item pool, and that a 'debiased' instrument would result (p. 180).

The disconcertingly large number of uninterpretable statistically biased

items leaves the test maker with a dilemma. Has the statistical indicator uncovered a real instance of bias, revealing a blind spot in the conceptualization of the test construct, or is the large bias index a statistical artifact, that is, not a valid sign of bias (see Shepard, 1981)? We are aware of the potential for artifactual errors in the bias methods. These artifactual explanations become all the more plausible when the bias results seem uninterpretable.

Control of Statistical Artifacts

There are both random and systematic sources of error associated with IRT bias indices. For example, because the current statistical theory for maximum likelihood estimation in item response theory is only approximate, conclusions regarding group differences may by sample dependent (Boulson and Lissak, 1981; Lord, 1980). Lord (1980), in fact, proposed that replication or reliability studies should be carried out on independent but randomly-equivalent groups of blacks and whites. One purpose of the present research is to conduct such stability comparisons.

There is also some "art" involved in the implimentation of IRT procedures. Choices made in running computer programs to arrive at maximum likelihood estimates can have small but important effects. In the use of IRT specifically for studying item bias, a difficult stage in the procedures is the equating phase. As we will see, parameters must be estimated separately for two groups but then equated to the same scale for comparative purposes. Errors in the equating can produce spurious instances of bias. In simulation studies, for example, three-parameter

indices of bias only correlated on the order of .80 with generated bias (see Merz & Grossen, 1979; Rudner, Getson, & Knight, 1980). Since the three-parameter logistic model was used to create bias in these data initially, near perfect correlations might have been expected between simulated and detected bias. One must conclude that either sampling fluctuations or some implementation problem, as suggested above, prevented better "recovery" of the bias that had been built-in.

In addition to the replication or "cross-validation" method to control for unreliability, the degree of error in bias indices can also be assessed by means of baseline studies. Lord (1980) created random groups which he called "reds" and "blues" to check on the number of "significantly" biased items in a condition where there should be no bias. Similarly, Ironson and Subkoviak (1979) used white-white comparison groups to assess the validity of both classical and latent-trait bias indices. In this research we will use white-white comparison groups to study not only the amount of "bias" due only to sampling errors but also to establish numeric base line values for interpreting bias indices that lack distribution theory.

Artifactual problems associated with random sampling error will be exacerbated when the groups to be compared differ in mean ability on the test. Angoff (1982) suggested that the classical p-value method would be more valid for detecting bias (rather than confounding differential difficulty with item discrimination) if groups were equal or nearly equal on the trait initially. For example, we might expect fewer artifactual problems in most male-female comparisons than with black-white comparisons. Even the three-parameter indices, which are theoretically sample invariant,

may be unstable when differences between groups are large. We know, for example, that latent-trait equating procedures are more stable for horizontal equatings (different tests, same grade) than unstable for vertical equatings (same test, different grades). The vertical equating problem where groups are located in very different regions of the ability continuum, is analogous to bias studies where groups have large mean group differences. Again referring to classical methods, Angoff (1982) suggested that an appropriate baseline for interpreting bias indices would not be just randomly equivalent white groups but white groups that differed in mean ability. This is the same analysis strategy used by Jensen (1974) when he created pseudo-ethnic groups, that is white groups selected on age (with a mean difference of two years) to simulate the average black-white differences. When we know that the statistical techniques are intended to correct for group differences but may do so imperfectly, the point is to simulate with all-white data what the effect might be of mean differences only. In the current research, pseudo-ethnic comparisons are used in addition to randomly-equivalent white groups.

Purpose Summary

The substantive purpose of this research is to study item bias between black and white examinees on both a mathematics and vocabulary test. The theoretically preferred three-parameter IRT approach will be used with optimal techniques for computing bias indices based on previous research. The major focus of the research is methodological rather than substantive. To assess the amount of artifactual (i.e., spurious) bias identified, both randomly equivalent white groups and extreme white groups

pseudo-ethnic comparisons) will be used. To identify the particular instances of unstable bias indices, cross-validations or replication analyses will be performed with randomly equivalent black and white groups. Finally, items found to be consistently biased will be inspected for substantive characteristics. It is hypothesized that once artifactual instances of bias are better controlled, the results should be more interpretable than has been the case with previous bias studies.

Method

Data Source

The data used for this investigation are from the High School and Beyond (HSB) data files available from the National Center for Education Statistics.[3] The HSB sample includes over 30,000 high school sophomores and 28,000 seniors, from a representative probability sample of the nation's tenth- and twelfth-grade populations. The test and questionnaire data were collected in the spring of 1980 by the National Opinion Research Center under contract with NCES. The particular examinees selected for study were black and white seniors. Unless otherwise specified (e.g., when pseudo-ethnic groups were created), the subsamples used were selected at random from the larger group of 3377 black or 17,928 white seniors (excluding Hispanics). The following

---

[3]We are grateful to Dr. Samuel Peng and Jeffrey Owings for providing a pupil file of individual item data as well as the publicly available aggregate tapes for pupils, parents, and schools.

study samples were created:[4]

Math Test

Comparison 1:	W1, B1	n = 1500 whites, 1500 blacks
Comparison 2:	W2, B2	n = 1500 whites, 1500 blacks
	(sampling for comparison 1 and comparison 2 was without replacement so the samples were independent.)	
Comparison 3:	W1, W2	(the white samples from comparisons 1 and 2)
Comparison 4:	B1, B2	(the black samples from comparisons 1 and 2)
Comparison 5:	W1, and Pseudo B (W3)	(the white sample from comparison 1 and white sample, n = 1500 selected to match the distribution of B1 on math total score)

Vocabulary Test

Comparison 1:	W4, B4	n = 1500 whites, 1500 blacks
Comparison 2:	W5, B5	n = 1500 whites, 1500 blacks
Comparison 3:	W4, W5	(the white samples from comparisons 4 and 5)

[4]In keeping with the constraints of LOGIST, examinees were excluded if their scores on the relevant test were 0% or 100%. Abilities (θ's) cannot be accurately estimated for subjects who are at the ceiling of the test; zero scores may not represent a valid administration and surely do not indicate that the examinee has been "measured."

The two tests analyzed were the senior Mathematics and Vocabulary tests. Although both tests were administered in two parts, we treated the combined item sets (32 in math and 27 on vocabulary) as single tests. More will be said about the nature of the items and the factor structure of the combined tests in the discussion of results. The math test was primarily a basic skills test involving simple operations, reading a graph, calculating a per unit cost, and comparing rates or distances. Four of the math items required some familiarity with basic algebra or geometry at a level that is usually included in K-8 curricula. The Vocabulary test was relatively more difficult; the average percent of seniors answering items correctly was .46 compared to .58 for math items. The content of the vocabulary test was clearly aimed at a higher grade level, either high school or in some cases college level. The words are almost exclusively Latin roots. According to word frequency counts in representative materials for grades 1 to 9 (Carroll, Davies & Richman, 1971) the vocabulary items are relatively unfamiliar. The standardized frequency indices indicate that the words are found either not at all in junior high level materials or at a rate of about one in a million or one in ten million words. Examples of words with similar frequencies would be: fanatic, recalcitrant, marauding, permeated, reciprocating, and crevasses.

IRT Bias Method

Item response theory permits the expression of examinee responses to individual test items as a function of the underlying ability or trait measured by the test. In the results section of the paper, there are

illustrations of these item characteristic curves (ICCs) or item response functions. The horizontal dimension of the graph is the ability (or θ) scale. For each item, a monotonic increasing curve reflects the probability of getting the item correct for increasing values of θ. The ICC is defined by three parameters: (1) the a parameter is proportional to the slope of the curve at the inflection point and represents the item's discrimination; (2) the b parameter reflects the item's difficulty and is a location on the θ ability dimension. When there is no guessing, b is the point where the probability of getting the item correct is 50%. (3) the c parameter is often referred to as the "pseudo-guessing" parameter. It is the lower asymptote of the curve and represents the probability of getting the item correct for examinees of extremely low ability.

The IRT method for detecting item bias is based on the comparison of item characteristic curves estimated separately for two groups. The ICCs reflect the probability of getting the item right as a function of ability. If an item is unbiased, examinees of equal ability should have equal probabilities of success on the item regardless of group membership; that is, the ICC's for different groups should be the same. If ICCs for two groups differ by more than sampling error, the item is apparently not measuring the same underlying trait for both groups (at least not to the same degree) and is therefore "biased."

It should be noted that IRT models rest on an assumption of unidimensionality, i.e., that the items in the test all assess the same underlying trait and that only ability on that trait, not some other trait, influences item performance. In the results section we present factor analyses as

supporting evidence that this assumption is met for these data. We did not devote much attention to prior testing of this assumption, however, because in a sense the bias studies themselves are addressed to the issue of unidimensionality. In fact, multidimensionality will be detected as bias by the IRT method so long as group differences are not uniform across the different traits. As stated by Linn, Levine, Hastings and Wardrop (1981), "Bias may generally be conceptualized as multidimensionality confounding differences on a primary trait with differences on a secondary trait" (p. 161).

The LOGIST program (Wood, & Lord, 1976; Wood, Wingersky & Lord, 1976) was used to estimate the person abilities and item parameters. Because the chance (c) parameters are difficult to estimate even with large sample sizes, we followed the technique suggested by Lord (1980) whereby c's were estimated in a combined analysis and then fixed at that value for the separate analyses within ethnic groups. This aggregate or composit analysis was done only at the level of each study comparison. That is, black and white samples chosen for separate replications were not combined to get even more stable estimates of c. Rather, we wished to preserve the separateness of each comparison study and do each as if it were the only data available to the researcher. Additional particular information about how the LOGIST program was implemented is given in the results section.

Scale Equating

Once item parameters (defining the ICCs) have been estimated separately for two ethnic groups, the ICCs for each item must be compared

to detect bias. However, because the θ scales from an IRT analysis are arbitrary (set with $\bar{X} = 0$ and $s = 1$ for the given sample), the ICCs from separate analyses are not directly comparable. The ICCs must first be equated to the same scale. To make the adjustment, we used a linear transformation of the b parameters as described in Linn, Levine, Hastings, and Wardrop (1980, Appendix B). Briefly, the equating is determined by a best fitting line which adjusts for the difference in average item b values and has a slope equal to the ratio of the standard deviations of the two sets of b's. In computing means and variances, b parameters were weighted by the inverse of the variance error in estimating b. Therefore, items with poorly estimated b's contributed least to the equating. Once the linear equation was obtained, the b parameters for the second sample were recomputed in the metric of the first group, e.g., in this case, the black parameters were converted to the white scale.

After the bs were adjusted, the same equating constants (the slope and intercept) were also used to transform the θ values. Finally, the a parameters were equated using the inverse of the slope determined for the b equating (Lord, 1980, p. 36). The c parameters do not require equating.

Bias Indices

For an individual test item, bias is defined as the difference in the probability of answering correctly, given equal ability. Once item characteristic curves have been adjusted to the same scale, differences in the probability of a correct response are synonymous with differences in the ICC's. Several different indices were used to quantify ICC

differences between groups.

Unsigned Indices

1. Unsigned Area (UA). As described in Shepard, Camilli and Averill (1981), the area between two ICC functions was evaluated as a definite integral for an item i:

$$\int_{-3}^{+3} \{\hat{P}_{iW}(\theta) - \hat{P}_{iB}(\theta)\}\, d\theta$$

2. Sum of Squares 1 (SOS1). Linn, Levine, Hastings and Wardrop (1980) developed both weighted and unweighted sums of squares statistics. The following index is similar but is "self weighting," in the sense that squared differences in probabilities are summed for every value of θ that occurs, rather than creating intervals on the θ scale and using the midpoint of each interval. Thus, probability differences in the region where the most data occurs will contribute more to the index.

$$SOS1_i = \frac{1}{n_W+n_B} \sum_{j=1}^{n_W+n_B} \{\hat{P}_{iW}(\theta_j) - \hat{P}_{iB}(\theta_j)\}^2$$

3. Sum of Squares 2 (SOS2). SOS2 is similar to SOS1 except that squared differences in probabilities were weighted by the inverse of the variance error of the difference in ICCs for each given value of θ.

$$SOS2_i = \frac{1}{n_W+n_B} \sum_{j=1}^{n_W+n_B} \frac{\{\hat{P}_{iW}(\theta_j) - \hat{P}_{iB}(\theta_j)\}^2}{\hat{\sigma}^2_{P_{iW}-P_{iB}}}$$

Formulae for computing the variance error of a point on an estimated ICC are given in Linn, Levine, Hastings and Wardrop (Appendix A, 1980). Following their reasoning, P differences contributed less to the weighted index if either P were poorly estimated.

4. Chi-square (χ^2). Lord (1980) proposed an asymptotic significance test to compare a and b differences between groups simultaneously. By the following chi-square formula, the hypothesis is tested that the vector of a and b differences is different from the vector (0, 0):

$$\chi_i^2 = V_i' \, \Sigma^{-1}_{ab_i} \, V_i$$

$$\text{where } V_i = \begin{Bmatrix} \hat{a}_{iW} - \hat{a}_{iB} \\ \hat{b}_{iW} - \hat{b}_{iB} \end{Bmatrix}$$

Signed Indices

All of the above unsigned indices reflect the magnitude of the differences between ICCs for two groups, but they do not carry signs to indicate the direction of the bias, i.e., which group has the lower probability of a correct answer. In fact when the item characteristic curves cross, one group is not consistently disadvantaged. Rather, one group is "ahead" in one region of the graph, but behind in another region.

Signed indices are computed similarly to the corresponding unsigned indices. When the ICCs do not cross, the absolute values of the indices are the same but with a sign attached to show the direction of bias. When the curves cross, "bias" in two regions of the curve may be offsetting.

5. Signed Area (SA). When the ICCs for two groups did not cross in the region from -3 to +3, the SA was equal to the UA except that a negative sign was attached if the item was biased against whites, i.e., if whites had a lower probability of getting the item right given Θ. If the ICCs did cross, Θ^* was found as the root of the equation $P_W(\Theta) = P_B(\Theta)$. Then the integral was evaluated from -3 to Θ^* and Θ^* to 3. The signed area was the difference between these two areas and carried the sign of the larger area.

6. Sum of Squares 3 (SOS3). SOS3 is the "signed sum of squares" index analogous to SOS1. By multiplying $\left[\hat{P}_{iW}(\Theta) - \hat{P}_{iB}(\Theta)\right]$ times its absolute value, rather than squaring the difference, the sign of the difference is preserved.

$$SOS3_i = \frac{1}{n_W+n_B} \sum_{j=1}^{n_W+n_B} \{\hat{P}_{iW}(\Theta_j) - \hat{P}_{iB}(\Theta_j)\} \; |\hat{P}_{iW}(\Theta_j) - \hat{P}_{iB}(\Theta_j)|$$

7. Sum of Squares 4 (SOS4). SOS4 is the weighted sum parallel to SOS2. It is computationally the same as SOS3 except that every squared difference is weighted by the inverse of the variance error of the difference.

$$SOS4_i = \frac{1}{n_W+n_B} \sum_{j=1}^{n_W+n_B} \frac{\{\hat{P}_{iW}(\theta_j) - \hat{P}_{iB}(\theta_j)\} \quad |\hat{P}_{iW}(\theta_j) - \hat{P}_{iB}(\theta_j)|}{\hat{\sigma}^2_{P_{iW}-P_{iB}}}$$

8. a and b differences (AD),(BD). Simple differences between a parameters ($a_W - a_B$) and b parameters ($b_W - b_B$)[5] were computed. These differences were of interest because of their relation to other bias indices. The a and b differences were not interpreted as bias indices themselves since separately they do not characterize ICC differences well. As Linn et al., (1980) have shown, a and b parameters could be substantially different for two groups but not result in any practical differences in the ICCs. This would be true, for example, if an item were extremely difficult in both groups so the b's diverged in a θ region where few examinees existed.

---

[5] Note that the black b was always subtracted from the white b, since this corresponds to the order of subtraction in all the other indices. However, because a high b means the reverse of a high probability of correct response, the signs will have the opposite meaning.

Results and Discussion

Factor Analyses

Factor analyses were performed on the mathematics and vocabulary tests to determine if they could be considered unidimensional. Tetrachoric correlations were obtained using the total senior sample of 25,069. Principal factors were extracted after iterating for communalities. Each factor with an eigenvalue greater than one was retained for rotation. An oblique solution was obtained by direct oblimin transformation with $\Delta = 0$.

In the math test the first unrotated factor accounted for 30% of the total variance. Four additional factors that met the minimum eigenvalue criterion of one accounted for 5%, 4%, and 3% of the variance, respectively. Similarly, on the vocabulary test the first unrotated factor comprised 30% of the total variance. In this case, there were only two additional factors, accounting for 6% and 4% of the variance, respectively. For both tests we interpreted the results as reasonably strong evidence of unidimensionality. First, the percentage of total variance explained by the first factor exceeds Reckase's (1979) minimum of 20% needed to assure stable item parameters. Also, an inspection of the scree plot of latent roots suggested that only the first eigenvalues deviated from the gradual rise that could be expected from factoring uncorrelated variables.

As stated previously, the singularity of the trait measured by the test for the population generally may not be measured so well in particular subgroups. It is the purpose of the bias analysis, in fact, to test whether

this unidimensionality is true for both black and white groups.

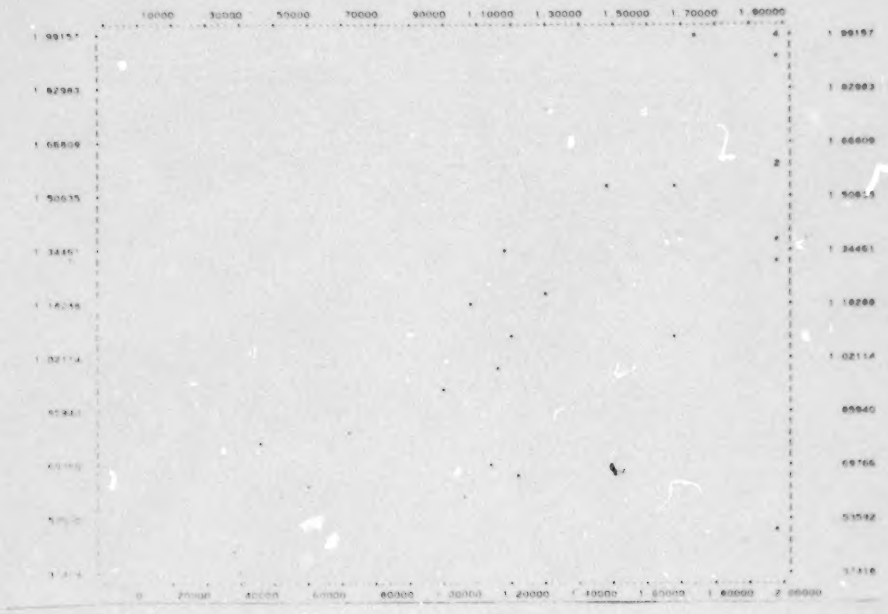
Equating and an Overall Picture of Bias

Equating of a and b parameters from separate black and white analyses is a procedural step leading to direct comparisons of the ICC's for the two groups. The equating results also are useful themselves to provide an overall picture of how similar the test is for blacks and whites.

In Figure 1b a scattergram for 29 of the 32 math items depicts the degree of relationship between the b parameters obtained for white sample 1 (W1) and black sample 1 (B1). (Items were excluded from the equating if either one parameter or the standard error of a parameter could not be estimated in either group). The correlation between b parameters in the first comparison was .97. A very similar graph and identical correlation, r =.97, were obtained for the W2,B2 comparison. It is important, however, to contrast Figure 1b with Figure 2b. Figure 2 shows the equating for the white-white comparison study (W1,W2)(r_b=.996). In the second scatterplot the b values differ only by sampling error. In contrast, the greater dispersion in the black-white comparison suggests that for these groups some of the item b's are more different than can be attributed to sampling fluctuations, i.e., there is apparently some bias in the test.

Figures 1a and 2a are the corresponding scattergrams for a parameters from the W1,B1 and W1,W2 analyses. In the first comparison the white-black a parameters were correlated .78. In the parallel sample 2 white-black comparison (not shown), the a correlation was .81. These correlations, obtained under conditions with some bias present, are in contrast to the white-white correlation in Figure 2a of .95. It should be

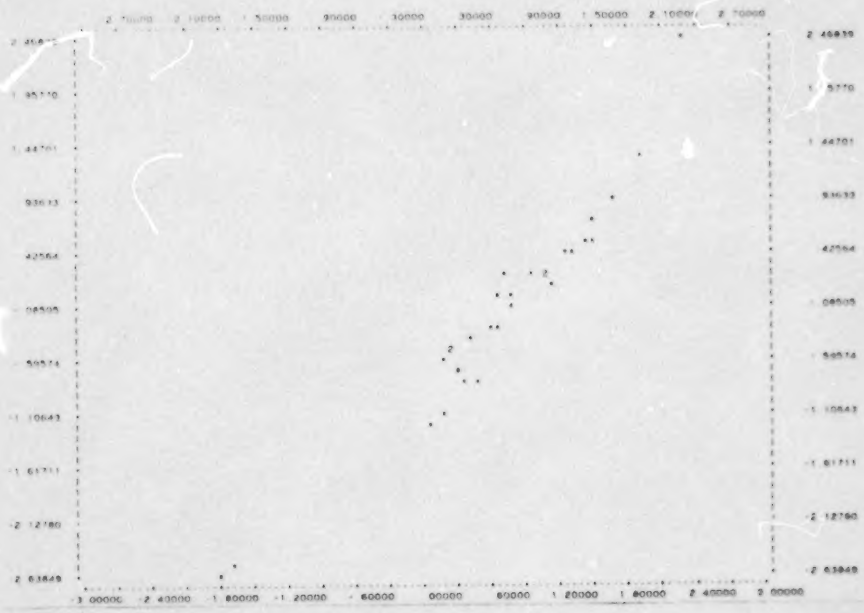
Figure 1a: Scattergram of a parameters (W1,B1)



white a's

black a's

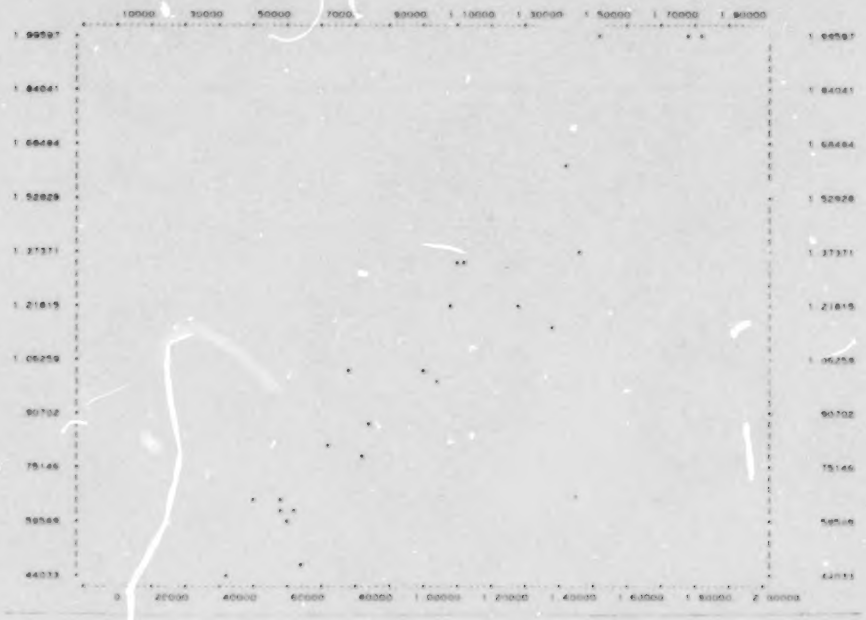
Figure 1b: Scattergram of b parameters (W1,B1)



white b's

black b's

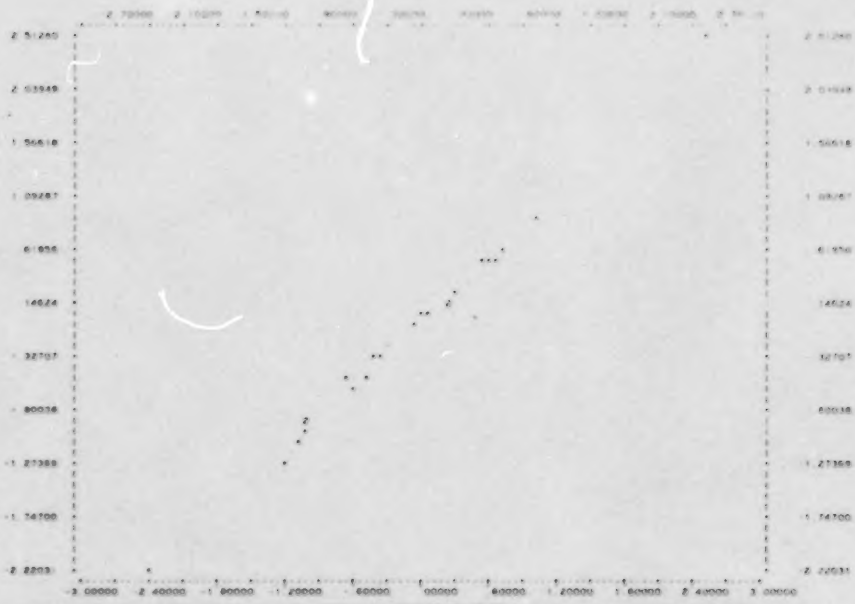
Figure 2a: Scattergram of a parameters (W1,W2)



white sample 1 a's

white sample 2 a's

Figure 2b: Scattergram of b parameters (W1,W2)



white sample 1 b's

white sample 2 b's

noted that even in the white-white analysis the equivalence of the a's was not as good as for the b parameters. This illustrates the relative instability of the a estimates (and explains why we prefer to use a function of the b equating to transform the a values). In the case of the black-white comparison the modest a correlation may also suggest that a highly correlated but slightly different trait is measured in the white and black groups.

For the methodological purpose of this research it is especially significant that there is apparently some bias between the measurement of blacks and whites on the math test. The equivalence of the bias results across parallel comparisons, the intercorrelation of bias indices, and the comparison to a white-white baseline are all informative only if there is bias to be detected. There is little to be learned about the consistency or validity of the bias indices if they are applied in circumstances of no bias.

The equating results for the Vocabulary test followed the same general pattern as that found on the math test. Overall, the adequacy of estimation was not quite so good for the vocabulary test, and hence the accuracy of the equating suffered. Even in the white-white analysis the b's were correlated .99 but the a's only .83 (compared to .95 for a's on the math test). For the two white-black comparisons the b's were correlated .95 and .94 respectively; a's from the two groups were correlated .80 and .56 respectively. The relatively poor correspondence between a parameters in study 2 compared to study 1, suggests the presence of estimation errors that are likely to turn up as artifactual (or at least

unstable) instances of bias.

Bias Indices with Replication for the Math Test

In Table 1, the bias indices are reported for mathematics items across five comparisons. To simplify the amount of information, a reduced set of indices is presented. The signed and unsigned areas and the asymptotic chi-square have been the most popular in the past. As we will see from the correlational results, the weighted and unweighted SOS statistics are highly similar, but there is some evidence for preferring the "behavior" of the weighted versions; therefore only SOS2 and SOS4 are shown.

The first two sets of indices, from Comparison 1 and Comparison 2, are the replicated bias studies based on randomly equivalent groups of blacks and whites. As will be discussed in the next section, a baseline for judging the magnitude of the bias indices was obtained from the white-white analysis. Index values that exceeded the largest number occurring in the white-white analysis are starred as biased in Table 1.

There were a substantial number of items with ICC differences that replicated across studies. Out of the 29 math items, for which ICCs were estimated in both groups, 10 were consistently "biased." (Three of these items were biased in favor of blacks.) We said items were consistently biased if they exceeded the cut-off on four or five of the indices in both studies. It is worth noting that fully one-third of the test items appear to be deviant by this relatively stringent rule. When we caution that item-bias methods are internal methods and hence unable to detect constant bias, this does not imply that we are limited to finding only one or two

Table 1

Signed and Unsigned Bias Indices for Math Items in Five Comparison Studies

	Comparison 1: W1, B1					Comparison 2: W2, B2					Content Classification: verbal or numeric	Comparison 3: W1, W2					Comparison 4: B1, B2					Comparison 5: W1, Pseudo B (W3)				
	Unsigned			Signed		Unsigned			Signed			Unsigned			Signed		Unsigned			Signed		Unsigned			Signed	
	UA	SOS2	χ^2	SA	SOS4	UA	SOS2	χ^2	SA	SOS4		UA	SOS2	χ^2	SA	SOS4	UA	SOS2	χ^2	SA	SOS4	UA	SOS2	χ^2	SA	SOS4
ITEM 1	.22*	4.06	49.12*	.08	-4.76	.05	3.73	6.40*	-.01	-2.72	n	.03	.04	.15	.03	-.04	.24*	38.49*	20.53*	-.03	38.31*	.05	1.13	2.16	-.05	-1.13
2	.13	2.05	3.76	.08	2.04	.18	3.28	4.73	.01	3.07	v+n	.05	.33	.34	.00	-.33	.13	.71	2.03	-.13	-.71	.15	1.26	3.39	.15	1.26
3	.33*	12.46*	20.93*	.33*	12.46*	.46*	13.53*	21.63*	.29*	18.72*	v	.17	.70	2.45	.12	-.33	.17	.91	1.81	.00	-.63	.31*	2.79	11.35*	.26*	2.30
4	.24*	6.81*	10.94*	.24*	6.81*	.33*	22.20*	14.11*	.23*	22.15*	n	.01	.02	.04	-.01	-.02	.08	.34	.69	-.07	-.34	.13	2.99	3.68	.11	2.97
5	.03	.96	.95	-.01	.78	.13	2.82	9.24*	.13	2.82	n	.09	1.06	4.71	-.08	.06	.03	.92	2.04	.03	.92	.09	1.31	12.07*	-.09	-1.26
6	.42*	15.59*	17.63*	.12	14.76*	.53*	18.71*	20.30*	.29*	18.25*	v	.07	.19	.71	.07	.19	.15	1.38	3.34	.14	1.38	.10	1.57	1.21	.03	1.53
7	.25*	9.81*	19.81*	.25*	9.81*	.23*	17.33*	22.50*	.23*	17.33*	v	.11	3.75	2.67	.03	-3.68	.08	.75	1.78	-.01	.22	.12	1.03	4.29	.12	1.03
8											n						.20	.59	3.65	.18*	-.35					
9	.19	1.34	5.69*	-.14	-.95	.43*	8.78*	16.74*	-.10	5.48	v	.12	1.25	2.14	.02	-1.05	.10	.52	1.31	.04	.41	.03	.13	.20	.03	.13
10	.14	4.35	6.66*	.01	-1.40	.14	2.81	4.81	.01	-.25	n	.04	.19	.48	-.01	.01						.08	1.22	2.40	.07	1.22
11	.31*	12.00*	14.87*	.10	11.60*	.51*	24.46*	19.25*	.05	22.43*	v	.06	.40	.38	-.01	.38	.08	2.56	1.01	-.03	2.56	.14	3.68	5.04	.12	3.68
12	.21	28.29*	23.20*	-.11	-27.89*	.26*	40.65*	42.45*	-.22*	-40.60*	n	.06	.47	.72	.02	-.42	.06	1.64	2.96	-.06	-1.64	.13	3.51	3.47	.03	-2.92
13	.13	4.26	8.16*	.05	-1.68						n	.05	1.41	1.90	.05	1.41	.05	1.36	2.65	-.05	-1.36	.09	1.10	3.36	.06	.24
14	.46*	10.08*	33.55*	.29*	.53	.50*	8.53*	33.69*	.34*	5.50	v	.13	1.94	2.47	.02	1.91	.11	1.21	3.37	.05	-.52	.16	2.06	5.68*	.16	2.06
15	.15	.48	2.27	.14	.40	.18	1.33	3.15	.13	-.70	v+n	.12	2.35	4.22	-.04	1.89	.03	.55	.17	.00	-.55	.19	1.11	2.89	.13	-.40
16	.23*	17.57*	42.44*	-.23*	-17.57*	.19	8.55*	20.88*	-.19*	-8.55	n	.06	1.37	2.32	-.04	-1.27						.10	2.99	7.79*	-.10	-2.99
17	.23*	12.17*	23.38*	-.22*	-12.13*	.34*	21.66*	34.56*	-.27*	-21.28	n	.10	3.74	3.70	.07	3.63	.05	.26	1.07	.05	.26	.16	6.03	6.46*	.04	4.49
18	.38*	6.21	13.40*	-.09	-.27	.27*	3.05	7.29*	-.18*	2.64	v	.09	.79	1.03	.03	.76	.07	1.78	1.49	-.02	-1.78	.32*	5.46	9.18*	.04	4.25
19	.01	.01	.02	.00	.00	.02	.03	.08	-.02	-.03	n	.06	.80	1.25	-.06	-.80	.05	.28	.87	-.05	-.28	.15	2.53	4.13	.13	2.51
20	.14	3.83	7.08*	.14	3.83	.15	2.86	7.97*	.15	2.86	v	.06	1.21	1.88	-.06	-1.21	.08	.66	2.20	-.05	.32	.11	2.92	3.46	.01	-2.11
21	.08	3.14	5.11	.07	3.10	.07	1.74	2.42	.05	1.68	n	.03	.47	.71	.03	.47	.05	4.66	1.78	-.01	4.64	.06	1.20	2.49	.06	1.20
22	.23*	13.79*	29.20*	.23*	13.78*	.25*	13.26*	25.73*	.25*	13.26*	v	.04	.28	.92	.02	.23	.03	.57	.28	.02	-.56	.05	.71	1.47	-.04	-.64
23											v											.38*	2.41	5.25	.38*	2.41
24	.06	.90	1.15	.03	.88	.04	.22	.33	.03	.22	v	.10	.42	1.77	.05	-.20	.07	.87	.72	.02	-.86	.15	1.56	2.75	-.09	.11
25	.06	.82	.93	-.06	-.82	.19	.15	2.04	.19*	.12	n															
26	.06	.37	1.04	.04	.12	.08	1.63	3.06	-.08	-1.63	n	.09	2.19	3.15	.02	1.74	.13	3.59	3.95	-.08	3.50	.05	.27	.85	.05	.26
27	.08	.28	1.73	.08	.19	.13	.47	3.09	.12	.28	n	.01	.01	.12	-.01	-.01	.03	.62	.53	.03	.62	.09	3.98	3.61	.02	-3.73
28	.08	1.50	3.90	-.08	-1.50	.06	2.20	1.83	-.03	1.97	n						.11	6.31	11.54*	.06	5.61	.07	.94	1.83	-.05	.65
29	.13	.77	2.73	-.13	-.77	.15	2.32	2.89	.06	2.31	n	.17	.56	4.57	-.17*	-.56	.26*	.43	4.21	-.16	.23	.01	.01	.02	.00	.01
30	.68*	20.07*	8.09*	-.26*	19.87*	.10	.20	.47	.06	-.20	v	.23*	6.59*	5.37*	.07	6.59*	.23*	1.97	2.03	.08	-1.96					
31	.04	.23	.31	.03	-.20	.05	.69	.50	.02	-.65	n											.03	.20	.38	.02	-.18
32	.06	1.29	.85	.00	-1.28	.09	.70	.56	.06	-.67	n	.01	.12	.08	-.01	.12	.14	.72	1.56	.13	-.72	.15	12.20*	1.46	.01	12.10*

Note: To establish a baseline for judging the magnitude of the bias indices, the largest value for each index from the white-white comparison was used. Indices that exceeded this value in other comparisons are starred as "biased." For the sake of consistency the largest W1, W2 χ^2 of 5.37 was used, however 5.99 is the critical value for statistical significance at $\alpha = .05$.

discrepant items.

Figures 3-6 are item characteristic curves for blacks and whites on several illustrative math items. In Figure 3a, the graph for an unbiased item is shown. The two solid lines reflect the probability of a correct response, given θ, for blacks and whites. For all values of θ, the two groups have essentially equal probabilities of answering correctly. Figure 3b is an example of a biased item. The white curve is consistently above the black curve, so whites and blacks of equal ability do not have the same probability of success. (The curves for item 7 were similarily discrepant in comparison 2 with a slightly larger effect.)

The items in Figures 4a and 4b are also consistently biased in both comparisons. These graphs are more typical of most of the biased items in that the ICCs for the groups cross within the θ region of -2 to +2. Therefore, the bias in one region of the curve is partially offset by a reverse bias at the other end of the θ scale. Signed indices allow this cancelling effect to occur and therefore only show a large bias index if one group is overall more disadvantaged than the other. Even between the signed indices there is a difference in how bias is quantified. The signed area (SA) is a simple measure of the amount of squared difference between the curves. The signed SOS4 index is more heavily weighted in regions where more examinees are concentrated. In Figure 4a both the signed area and SOS4 index are large, whites have a considerable advantage over blacks for θ's above -1. In Figure 4b, the areas of advantage and disadvantage are more nearly equal, hence a near zero signed area. The SOS4 value for this same item is quite substantial, however, because more examinees of both groups, especially blacks, are located in the vicinity

Figure 3a: Item 21, Comparison 1 (W1,B1)
Unbiased in both comparisons

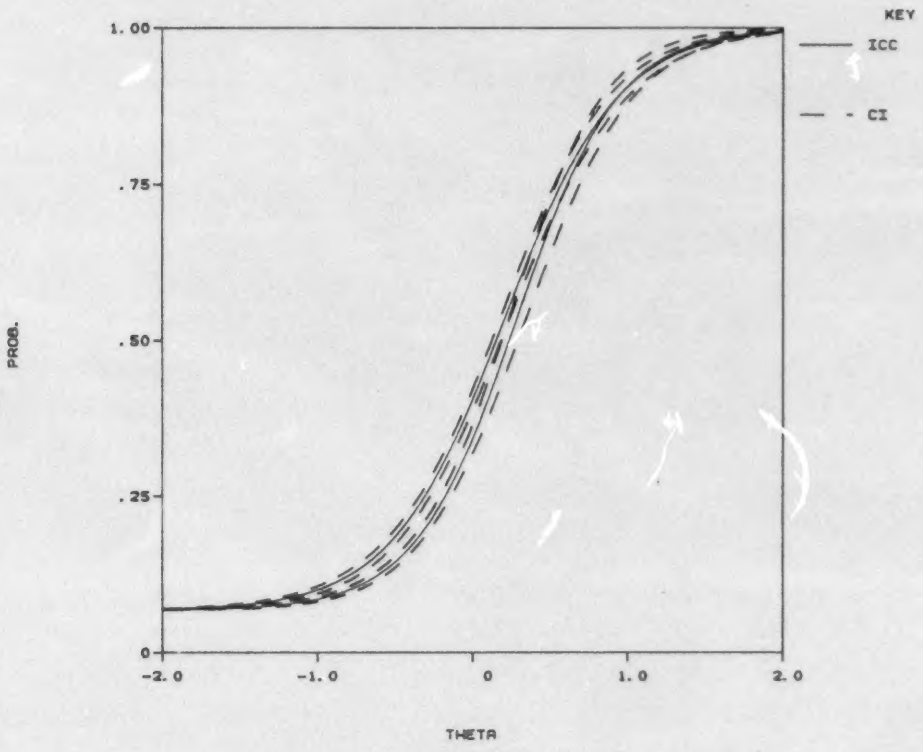


Figure 3b: Item 7, Comparison 1 (W1,B1)
Biased in both studies
ICCs do not cross, blacks have lower probability for all θ

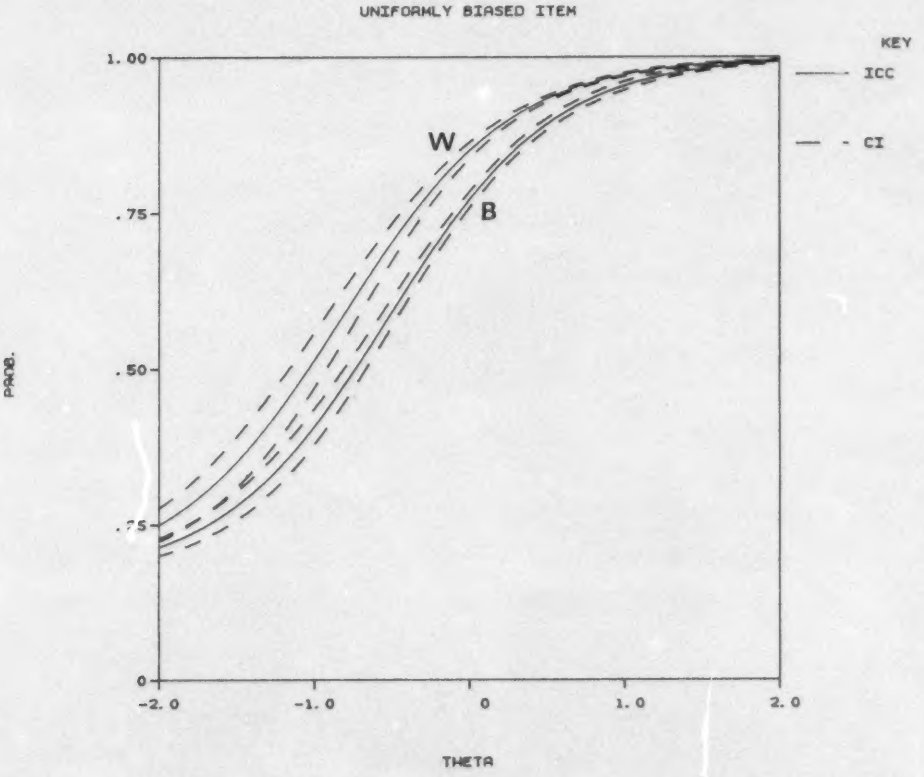


Figure 4a: Item 6, Comparison 2 (W2,B2)
Biased in both comparisons
ICCs cross, predominantly biased against blacks.

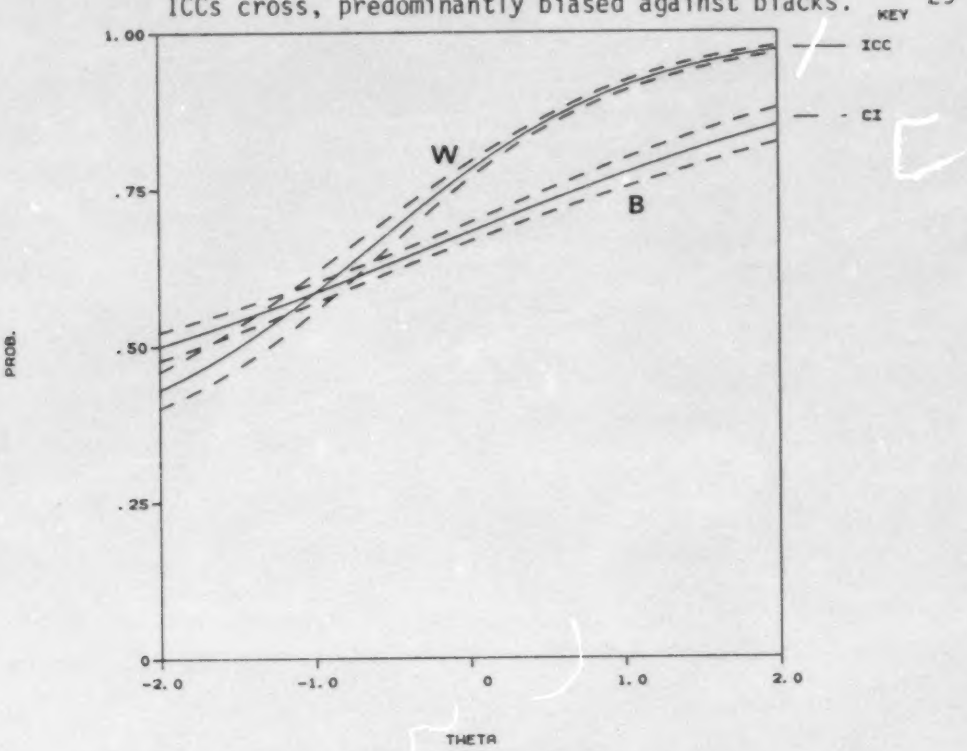
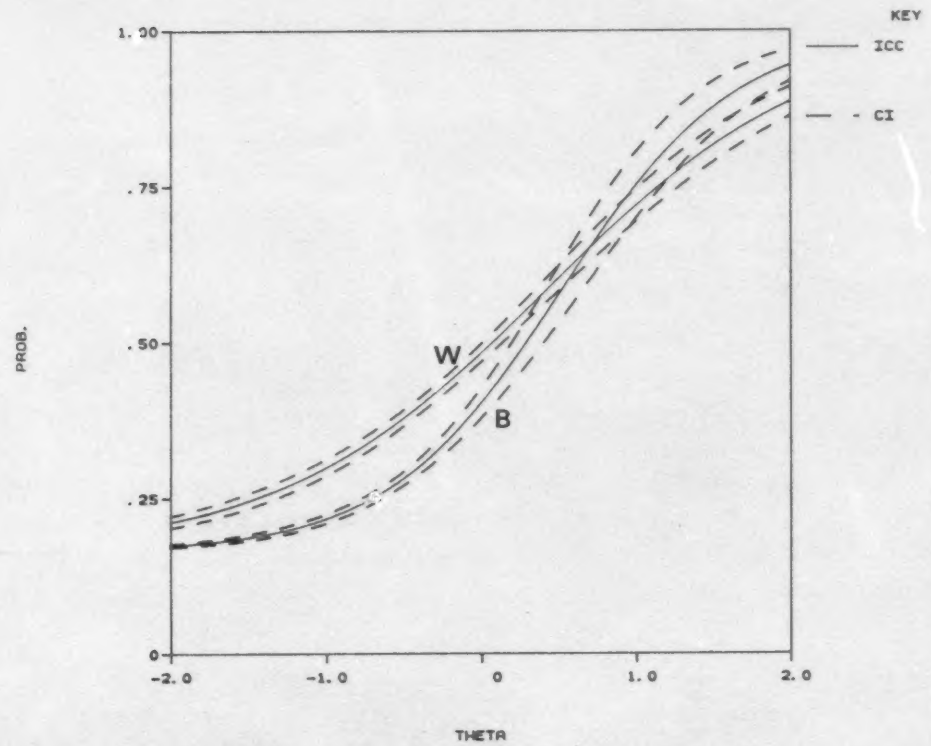


Figure 4b: Item 11, Comparison 1 (W1,B1)
Biased in both comparisons
ICCs cross, biased against blacks in
region where most blacks are located.



of -1 to 0 theta.

Two examples of items biased against whites are shown in Figures 5a and 5b, items 12 and 17 respectively. Two graphs from the parallel analyses are presented for item 17 to illustrate the replication results. The amount of similarity seen here between two independent but equivalent comparisons is fairly typical of the degree of stability found for consistently biased items (and for consistently unbiased items as well.)

Item 30, in Figure 6, is an "artifactually biased" item. All of the indices are substantially deviant in comparison 1 but not in comparison 2. Item 30 is very difficult for both groups. Hence, the a and b parameters must be estimated in a region where there is relatively little data. The difficulty in estimation is reflected in large standard errors. It should be noted however that even the statistics which take standard errors into account (χ^2, SOS2, SOS4), and the SOS measures which deemphasize discrepancies in regions with little data, had large values from this apparently spurious bias. Note also that Item 30 can be seen in the scattergram of b's (Figure 1b) as a clear outlier; misestimation of b's in both groups had a compound effect in comparison 1 that did not occur in comparison 2.

White-white and Black-black Comparisons on the Math Test

Item characteristic curves determined in two randomly equivalent groups should differ only by sampling error. Comparison 3 in Table 1 contains the bias results for two samples of whites (W1,W2). Logically, there should be no bias in this comparison and, indeed, inspection of these data indicates that all of the indices are appreciably smaller than in the white-black comparisons. Only item 30 which we know had estimation problems

Figure 5a: Item 12, Comparison 2 (W2,B2)
Biased against whites in both studies

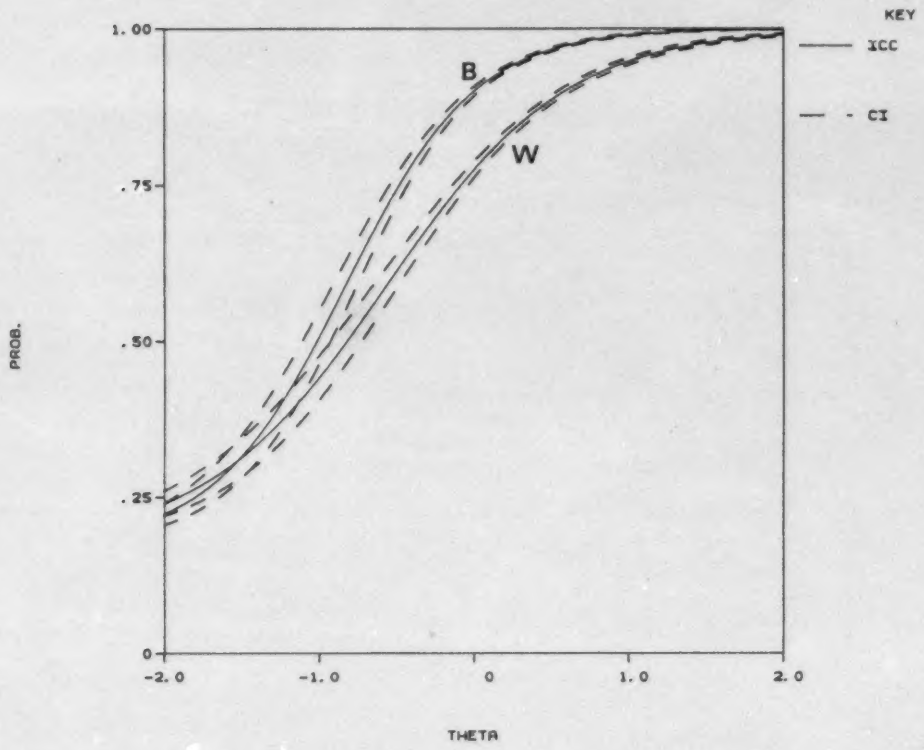


Figure 5b: Item 17, Comparison 1 and Comparison 2, respectively
Biased against whites in both studies (by different amounts).

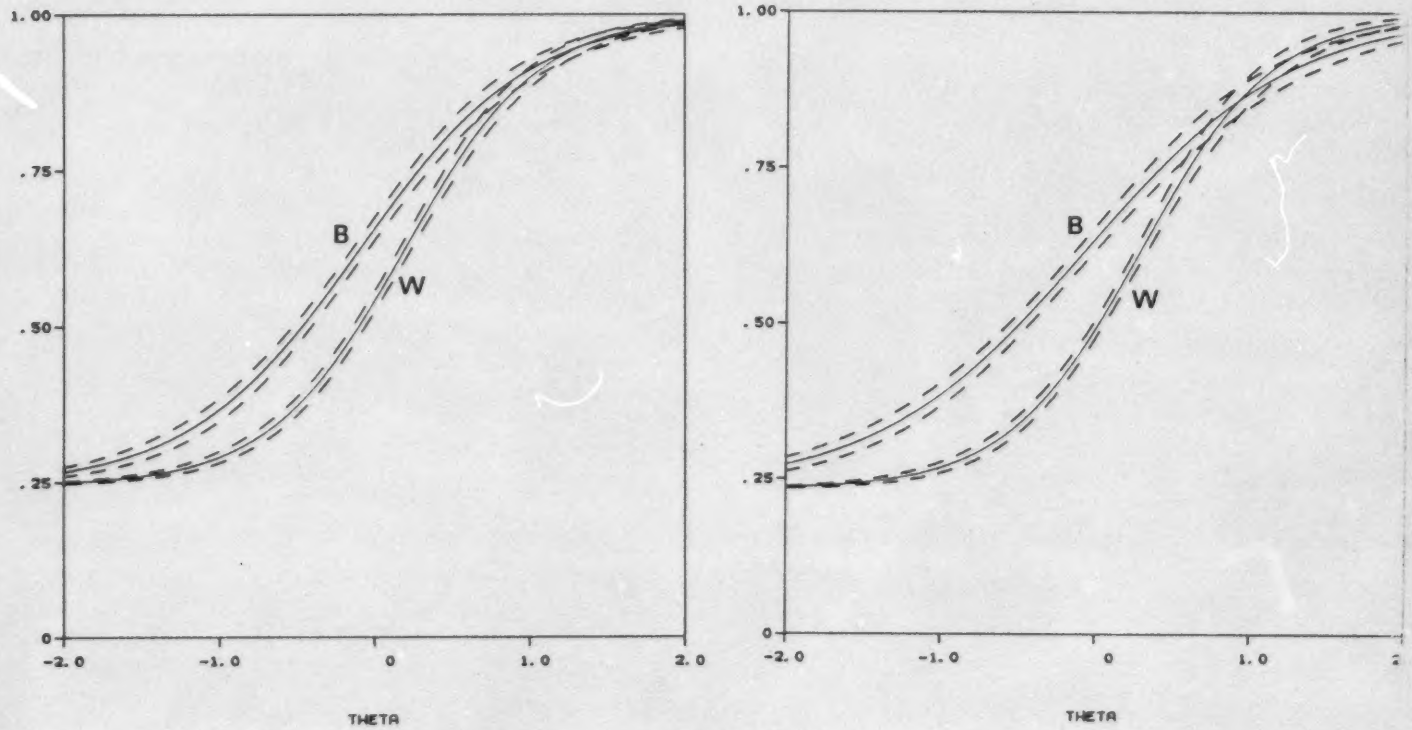


Figure 6a: Item 30, Comparison 1 (W1,B1)
Biased in comparison 1 only

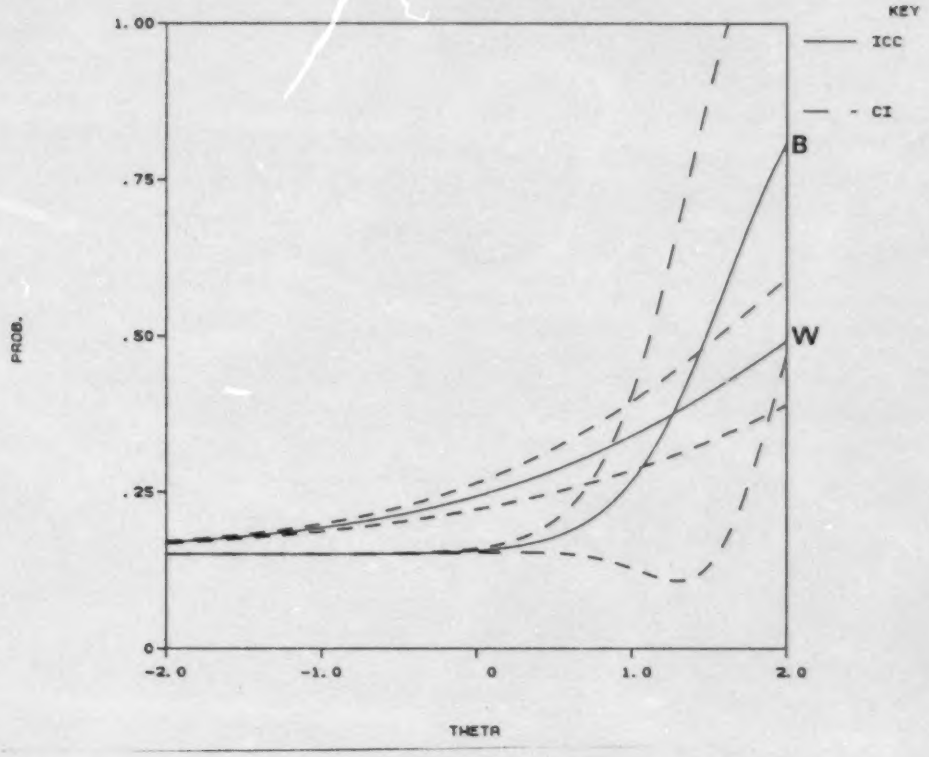
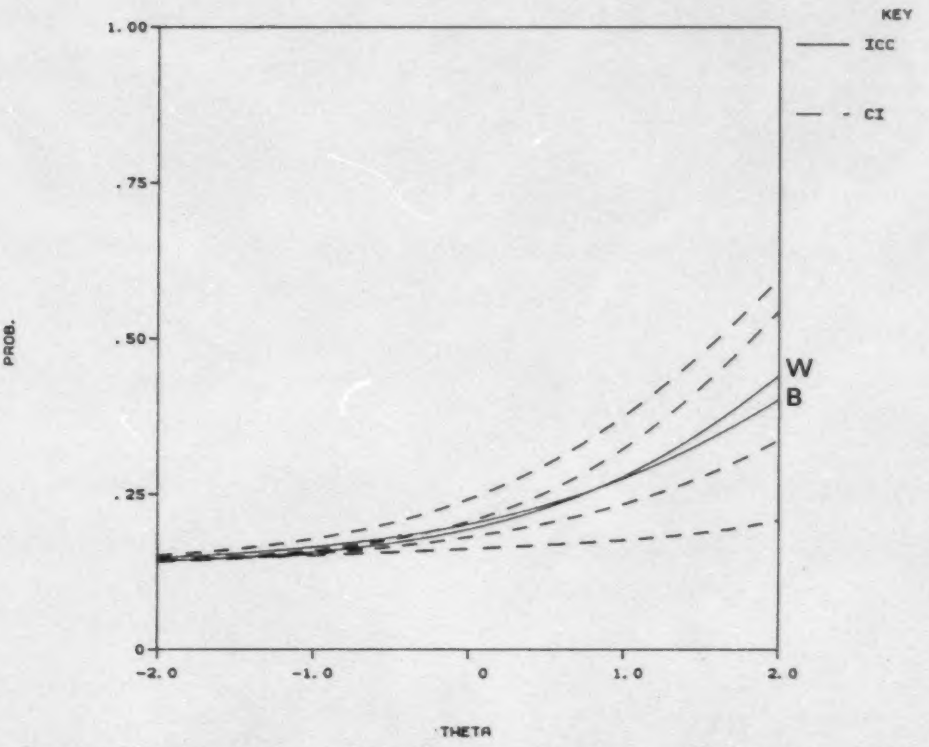


Figure 6b: Item 30, Comparison 2 (W2,B2)
Not biased in comparison 2



in sample 1, stands out with relatively large values. (Still, the numbers are much smaller than the corresponding indices in the between-ethnic comparison.)

The non-zero values of each index in Comparison 3 indicate the ranges in magnitude that occur as sampling fluctuations. Therefore we used the largest value of each index occurring in study 3 as the cut-off for evaluating bias in the black-white studies.

The stability of results in the two white samples was relatively dramatic. Therefore, we wondered if the white-white comparison would produce too stringent a baseline. It was conceivable that estimation problems could be more difficult in the black group. Although all samples were equal in size (n = 1500), something like a range restriction problem in the black group could make parameters more unstable for this group. This unreliability could then lead to spuriously large bias indices--especially if the more stable white-white analysis were used as the baseline.

To test the above hypothesis we also conducted a black-black "bias" analysis. Indeed, we did encounter more estimation problems than with previous analyses. All but two item ICCs had been estimated for sample B1 when c's were estimated in common with W1. However, standard errors could not be estimated for more than a third of the items when B1 was rerun with pooled c's from B2. We eventually were able to finesse the LOGIST runs by inputing initial item parameters from the B1, B2 run and by raising the upper limits on a's. After these estimation difficulties were resolved, however, ICCs comparisons for the two black samples (comparison 4, Table 1), did not result in a wholesale increase

in the number of large bias values. Therefore, we continued to use the baseline values obtained in the white-white study.

Pseudo-ethnic Comparison for the Math Test

It is conceivable that even IRT methods, which are theoretically sample invariant, may be inadequate when differences between groups are large. On the math total scores, blacks were .91σ below the white group. To what extent might the apparent item discrepancies in Table 1 be due to failure of the model to cope with mean differences in the separate ICC analyses? To answer this question, we created a pseudo-black sample. This group was selected at random from the original file of white examinees but with the probability of being selected constrained to match the relative frequency distribution of black total math scores. (We recognize the circularity implicit in matching on the very test to be analyzed; in a separate program of research we are using different sets of background variables external to the test, e.g., SES factors and instructional history, to study their effect on the issue of bias.) The white sample matched to the black distribution can give us a rough idea of the amount of deviance showing up in the bias indices solely as a function of mean group differences and sampling error. Because of regression effects on individual items however, the W1, Pseudo B(W3) comparison is not quite so extreme as the W1,B1 difference.

The results of the pseudo-ethnic bias study are shown as comparison 5 in Table 1. Note that there are very few large indices. Therefore, the large amount of deviance in the black-white analyses must be due to real differences in the functioning of the items across groups rather than

being due to artifacts of the mean difference in math achievement.

The chi-square index produced the greatest number of large values in the pseudo-ethnic comparison. For the four items where the χ^2 is starred as biased but no other index exceeded its cutoff, there was in each case a fairly big shift in the b parameter. As Linn et al., (1980) point out, it is questionable whether differences only in b parameters should be taken as evidence of bias. For these items the b shift is not reflected in overall ICC differences, or else the other indices would have shown large effects as well.

Correlations and Agreements Among Bias Indices

In subsequent sections the bias analyses for the Vocabulary Test will be presented and the nature and importance of the apparent bias in the Math Test will be explored. Here, we wish to discuss some methodological issues regarding the functioning of the bias statistics. Results are presented for both tests to check on the generalizability of study findings.

To examine the relationships between indices, within-study correlations were obtained for each comparison on each test. Tables 2 and 3 contain the within-comparison coefficients for the Math and Vocabulary Tests, respectively. As we explained in previous work (Shepard, Camilli, & Averill, 1981), Spearman rank-order correlations are preferred. With the Pearson r, one very extreme item will occasionally inflate or obscure the degree of relationship. When studying bias, congruence in the identification of extreme items is of primary interest; therefore, we did not wish to trim the distribution or eliminate outliers.

In both Table 2 and Table 3 the first two entries are for comparisons where some bias is present, i.e., these are the between-ethnic

Table

Intercorrelation of Bias Indices within Comparison
on the Math Test
(repeated for five comparisons)

Order of r's: B1,W1 n's = 30 items
B2,W2 29
W1,W2 27
B1,B2 27
W1, Pseudo B 29

	UA	SOS1	SOS2	χ^2	SA	SOS3	SOS4	AD	BD
		.89	.83	.84	.26	.42	.32	-.40	-.22
UA		.92	.78	.81	.28	.31	.42	.05	-.02
		.74	.73	.86	.26	.18	.07	-.28	.12
		.59	.23	.65	.10	.04	.05	.22	-.08
		.83	.65	.70	.55	.65	.51	-.27	-.46
			.90	.91	.20	.34	.19	-.33	-.18
			.81	.91	.24	.31	.32	.00	-.03
SOS1			.87	.94	.04	.17	.21	-.13	.32
			.53	.92	.01	-.03	.12	-.02	-.02
			.71	.79	.34	.54	.39	-.23	-.26
				.88	.08	.23	.17	-.40	-.04
				.90	.04	.13	.33	-.16	.12
SOS2				.87	.17	.23	.21	-.33	.20
				.50	.03	.06	.21	-.10	-.03
				.65	.14	.34	.36	-.36	-.04
					.22	.29	.02	-.44	-.24
					.14	.19	.23	-.16	-.03
χ^2					.07	.12	.16	-.16	.29
					-.04	.07	.26	-.10	.06
					.32	.34	.21	-.18	-.26
						.87	.59	-.02	-.92
						.96	.69	.11	-.74
SA						.73	.35	-.32	-.73
						.72	.01	.20	-.99
						.88	.57	.09	-.95
							.79	-.20	-.82
							.77	-.02	-.70
SOS3							.81	-.41	-.38
							.57	-.13	-.65
							.74	-.04	-.84
								-.01	-.44
								.00	-.28
SOS4								-.22	.01
								-.55	.05
								-.22	-.46
									.11
									-.02
AD									.19
									-.17
									-.11

Table 3

Intercorrelation of Bias Indices Within Comparison
on the Vocabulary Test
(repeated for three comparisons)

Order of r's: WA, B4
W5, B5
W4, W5

n = 25 items

	UA	SOS1	SOS2	χ^2	SA	SOS3	SOS4	AD	BD
		.86	.59	.71	.28	.38	.27	.42	-.31
UA		.88	.82	.85	-.04	.11	.35	-.14	.13
		.91	.64	.95	-.01	.11	.06	-.16	.01
			.71	.81	.21	.31	.17	.45	-.27
SOS1			.93	.94	-.05	.09	.17	.13	.04
			.78	.97	.01	.18	.24	-.05	-.05
				.95	-.04	.10	.15	.33	.06
SOS2				.88	-.04	.13	.34	-.02	.10
				.76	.24	.40	.36	.10	-.24
					.06	.19	.16	.47	-.04
χ^2					-.15	.02	.23	.21	.08
					.04	.17	.16	-.08	-.07
						.96	.57	.21	-.96
SA						.90	.49	-.10	-.90
						.85	.12	.25	-.97
							.75	.21	-.91
SOS3							.65	-.16	-.84
							.44	.10	-.81
								-.11	-.47
SOS4								-.33	-.35
								-.13	-.07
									-.18
AD									-.12
									-.20

comparisons. The remaining entries reflect the degree of correspondence between indices within a comparison where there are different amounts of sampling instability but presumably no bias. Although one might expect the correlations between indices to be higher in the presence of bias, this is in fact not the case. The indices which are highly similar to each other are similar whether they are quantifying extreme deviance or only sampling perturbations. Afterall, whatever these sampling fluctions are, they are constant within a given comparison.

The unsigned indices are highly correlated suggesting they will yield fairly redundant information. The signed indices are also correlated with each other. However, the SOS4 statistic and the other signed statistics are less highly intercorrelated than the unsigned indices.

It was on the basis of these within-study correlations that we eliminated the simple sum-of-squares statistics from some of the results tables. The SOS1 index is essentially redundant with both the unsigned area and SOS2; SOS3 gives nearly the same picture of bias as the signed area. Note also that the pattern of relationships among indices (across comparisons of different types) was highly similar for both the math and vocabulary tests.

The more important test of the stability and validity of the indices as signs of bias is the pattern of correlations between study comparisons. As indicated above, the within study correlations show consistency from both true and error sources of variance. The between comparison correlations are given in Tables 4 and 5 for the math and vocabulary test respectively. Again, rank-order correlations were computed (but Pearson r's

Table

Correlations of Each Bias Index with Itself
Across Study Comparisons on the Math Test

Spearman r
(Pearson r)

	UA	SOS1	SOS2	χ^2	SA	S)S3	SOS4	AD	BD
W1,B1 with W2,B2	.71 (.57)	.70 (.66)	.72 (.75)	.80 (.65)	.72 (.78)	.83 (.89)	.73 (.84)	.83 (.61)	.65 (.50)
W1,B1 with W1,W2	.33 (.53)	.15 (.32)	.08 (.29)	-.02 (-.14)	.08 (.19)	.27 (.05)	.12 (.17)	-.20 (.00)	.22 (.53)
W2,B2 with W1,W2	.27 (.17)	.06 (.02)	-.08 (-.05)	.08 (-.02)	.01 (.07)	.06 (-.18)	-.15 (-.21)	-.46 (-.43)	.08 (.39)
W1,B1 with B1,B2	.32 (.39)	.11 (-.04)	.26 (-.06)	.17 (.50)	-.11 (-.08)	.10 (.08)	-.05 (-.13)	-.21 (-.45)	-.15 (-.25)
W1,B1 with W1, Pseudo B	.32 (.44)	.18 (.20)	.36 (.15)	.33 (.25)	.49 (.55)	.22 (.30)	.18 (.23)	.28 (.05)	.41 (.29)
W1,W2 with W1, Pseudo B	.24 (.27)	.22 (.26)	-.24 (-.07)	.19 (.27)	.17 (.30)	.44 (.46)	.37 (.23)	.16 (.07)	.21 (.36)

Note: Only for the correlations between W1,B1 with W2,B2 is there the possibility for agreement when bias is present. For other correlations, one or both of the comparisons involved randomly equivalent groups or two white groups; therefore, there should be no consistent bias. These latter pairs do share some consistent errors, however, since in each case one of the samples is repeated in both comparisons. Only in the correlations below should there be both no bias and no sample redundancy.

W1,W2 with B1,B2	.32 (.47)	.27 (.13)	-.02 (-.15)	.21 (-.14)	.42 (.47)	.04 (-.14)	-.09 (-.08)	-.04 (-.01)	.09 (.23)

Table 5

Correlations of Each Bias Index with Itself Across Study Comparisons on the Vocabulary Test

Spearman r
(Pearson r)

	Indices								
	UA	SOS1	SOS2	χ^2	SA	SOS3	SOS4	AD	BD
W4,B4 with W5,B5	.60	.64	.86	.83	.63	.82	.84	.32	.56
	(.80)	(.89)	(.71)	(.80)	(.74)	(.91)	(.82)	(.49)	(.81)
W4,B4 with W4,W5	.60	.46	.18	.45	.00	.02	.12	.32	-.14
	(.40)	(.10)	(.23)	(.18)	(.06)	(-.08)	(.23)	(.64)	(-.41)
W5,B5 with W4,W5	.61	.41	.24	.45	-.49	-.31	-.09	-.33)	-.50
	(.44)	(.17)	(.45)	(.31)	(-.49)	(-.34)	(-.50)	(-.37)	(-.37)

Note: Only for the correlations between W4,B4 with W5,B4 is there the possibility for agreement when bias is present. The later pairs do share some consistent errors, however, since in both cases one of the white samples is repeated in both comparisons.

are also given in parentheses). These coefficients reflect how highly a bias index correlates with itself across study comparisons; that is, how consistently does it rank the 29 math items studied?

In a sense these coefficients can be examined for convergent and discriminant validity as in a multitrait-multimethod matrix. The first line of each table is where we expect to see the effect of the trait on the magnitude of the correlations. The trait is, of course, "bias" in the test items or differential functioning of the items due to cultural background. Only in the first row are the correlations between two randomly equivalent ethnic comparisons. It is here that we would expect to see consistency in the detection of bias. Indeed the degree of relationship is quite good; r's for the math test range from .70 to .83, for the vocabulary test they are on the order of .60 to .86. (Note that a and b differences are presented to study their correspondence with other statistics, but they are not interpreted as indices of bias.)

The subsequent rows in the between-group matrices contain correlations where bias should not be the source of agreement. In all the remaining rows at least one or both of the comparisons were between equivalent groups (either both white or black). These correlations should show discriminant validity or the lack of method-specific correlations. These correlations should be near zero, confirming a lack of bias when none exists conceptually. However, it should be noted that these pairs of comparisons do share some consistent errors since one sample is repeated in both comparisons. For example, we expect the correlation between indices obtained in the W1,B1 study and those from the B1,B2 study to correlate zero. Bias can be present in the first study but not the second. The two comparisons do,

however, share the B1 sample. Therefore the two studies could have some consistent spurious "bias" based on sample characteristics. Only in the last row of the math data (Table 4) are there correlations between conditions where there should be both no bias and no consistent sampling error.

The discriminant coefficients show the reduced relationships necessary to support the validity of the bias indices. For example the SOS2 statistic is correlated .72 with itself when bias is present in both studies; it is correlated only .02 to .36 across studies where bias is not the source of relationship. However, the pattern of high-trait, low-method correlations is not so good for the unsigned indices on the vocabulary test. Two reasons should be kept in mind; the vocabulary test is less biased and as we explained in previous research (Shepard, Camilli, & Averill, 1981), it is more difficult to show ranking consistency with unsigned indices because they are one-tailed distributions. That is, unsigned indices have both items biased against blacks and those biased against whites in the same tail of the distribution making it more difficult to demonstrate consistency across parallel studies.

We are tentatively prepared to recommend the SOS2, SOS3 and SOS4 indices as the more valid indices of bias. Not only are these statistics the most consistent in detecting bias in the ethnic comparisons, they also intercorrelate the least in situations of no bias. A minor caveat is warranted, however, regarding the two weighted measures (SOS2 and SOS4). Because in our method of IRT estimation we fixed c's from a common analysis, we assumed that standard errors for c were zero in the weighted SOS formulae. To the extent that this assumption was erroneous, especially for very easy items, the same false assumption could add spurious agreement to the

between-study consistency. We judged this effect to be very slight. This problem could not, of course, explain the desirable drop-off in correlations in contrasts where bias should not be present. The SOS3 statistic was not affected by this assumption.

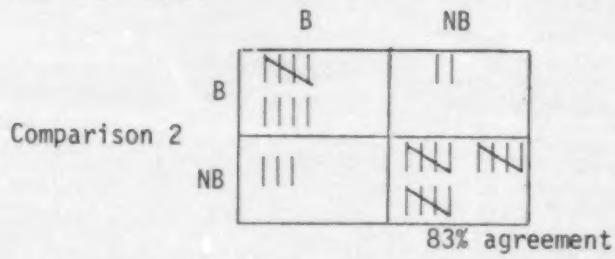
Correlation coefficients are only a crude method for summarizing the consistency of indices in identifying biased items. We are not actually interested in the consistency with which unbiased items are ranked. Rather, only the consistency at the extremes of the item rankings is important. Using the cut-offs determined from the white-white analysis, items were classified as either bias or not biased by each index. The contingency tables in Table 6 show the consistency of these dichotomous classifications from one black-white comparison to the other (on the Math Test). Here it should be clear that the SOS2 and SOS4 are relatively the best, and in an absolute sense, quite good at consistently classifying items as biased or not biased. The χ^2 statistic is next-best in the amount of replicated bias. But, as we explained earlier the χ^2 can consistently identify as biased items that have a large parameter difference but do not have a commensurate probability difference for most sampled θs. This occurs especially when items have large $\underline{b}$ differences at the extreme ranges of θ. The χ^2 index has the property of consistency but is less desirable on other grounds.

The agreement results found for the math test were only partially duplicated on the vocabulary test. The percentages of agreements were as follows: UA 70%; SOS1 75%; SOS2 90%; χ^2 85%; SA 70%; SOS3 55%; SOS4 75%. On the Vocabulary Test there was less bias; also on this test we had more difficulty justifying a particular cut-off from the white-white analysis.

Table 6

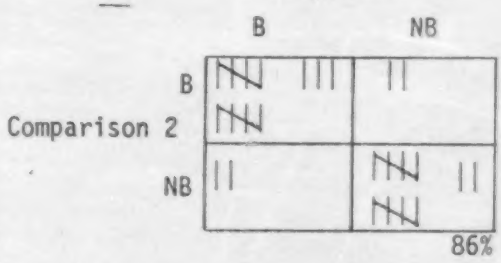
Agreement of Indices in Equivalent White-Black Comparisons on the Math Test

Unsigned Area



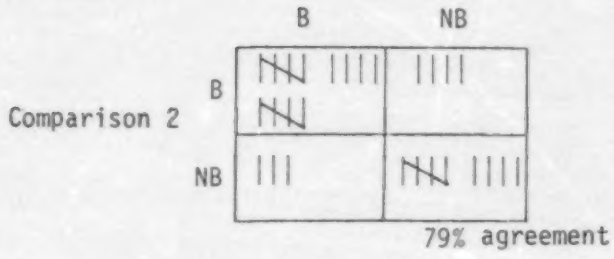
83% agreement

χ^2



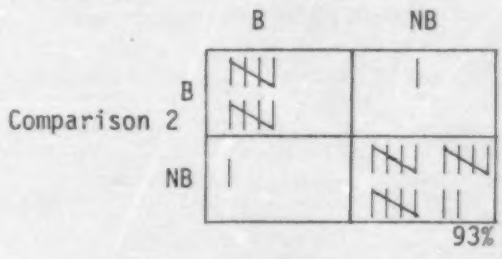
86% agreement

SOS1



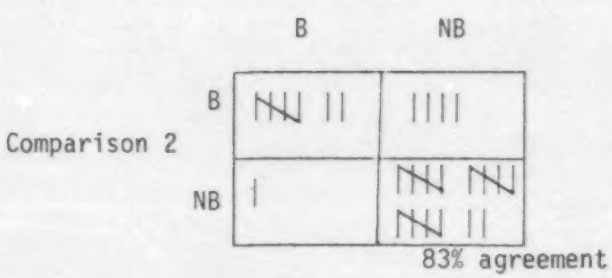
79% agreement

SOS2



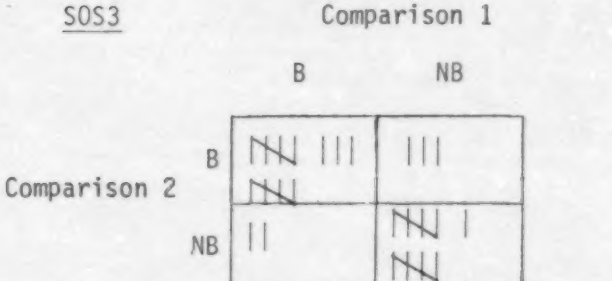
93% agreement

Signed Area



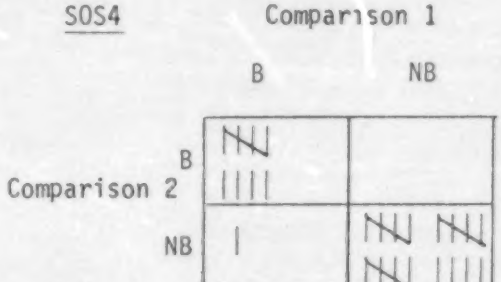
83% agreement

SOS3



83% agreement

SOS4



97% agreement

Note: These counts are based on the individual item data presented in Table 1. Biased items, starred in Table 1, had indices for a given comparison that exceeded the cut-off value determined from the white-white comparison. For the χ^2 index, however, the critical value of 5.99 was used here.

Substantive Interpretation of Bias Results for the Math Test

The original premise motivating this research was that the results of item bias analyses would be more interpretable if statistical artifacts could be controlled. Specifically, we expected to see more of a pattern in test items found to be biased if we studied only those items that were cross validated, i.e., found to be deviant in parallel black-white comparisons.

Once we had identified the consistently biased and unbiased items on the math test, we looked at the actual test questions. It was immediately obvious that the verbal math problem-solving was the source of the bias against blacks. (To be honest, the cross-validation did little to clarify this picture. The indices were consistent enough across studies that very nearly the same insight would have been gained by looking only at the results from one black-white comparison.)

All of the HSB math items (Part 1 and Part 2) had the following format:[6]

Directions: Each problem in this section consists of two quantities, one placed in Column A and one in Column B. You are to compare the two quantities and mark oval

A if the quantity in Column A is greater;
B if the quantity in Column B is greater;
C if the two quantities are equal;
D if the size relationship cannot be determined from the information given.

---

[6]We are grateful to Dr. Thomas Hilton and to the Educational Testing Service for providing us with copies of the test materials. We thank ETS and NCES for permission to reproduce the sample items and to create parallel item types to illustrate the nature of the verbal and numeric problems.

Sample Questions			Sample Answers
	Column A	Column B	
Example 1.	20 per cent of 10	10 per cent of 20	(A) (B) (C) (D)
Example 2.	6 × 6	12 + 12	(A) (B) (C) (D)

Answer C is marked in Example 1 since the quantity in Column A is equal to the quantity in Column B. Answer A is marked for Example 2 since the quantity in Column A is greater than the quantity in Column B.

We called items like Example 1 verbal and those like Example 2 numeric.

A more realistic illustration of the verbal-type items is provided by the two following questions. These items were written to parallel two actual test questions that were found to be consistently biased against blacks:

	Column A	Column B
1.	Number of centimeters between -7 cm and +8 cm	Number of centimeters between -8 cm and +7 cm
2.	Cost per pound at a rate of $4.00 for twenty pounds	Cost per pound at a rate of 3 pounds for 60¢.

A type of numeric problem found to be consistently biased in favor of blacks was parallel to the following example:

3.	326	$3(10)^3 + 2(10)^2 + 6(10)$

Numeric items that were consistently unbiased were similar to the following:

4. $\sqrt{16}$ 16

5. 5a 6x

The only numeric item found to be biased against blacks was comparable to this item:

33 ÷ 5 37 ÷ 5

If questions had a verbal phrase in one column and a numeral in the other column, we called them V + N. The classification of math items as verbal or numeric was shown in Table 1.

The following contingency table depicts the cross-tabulation of the bias results with item type:

Table 7

Bias Classification and Item Type for Math Items

	Verbal	V+N	Numeric
consistently biased against blacks	6		1
possibly biased against blacks	2		1
not biased	2	2	8
possibly biased against whites	1		
consistently biased against whites			3

These data show a striking degree of relationship suggesting that the bias indices are indeed sensitive to a change in meaning of the underlying trait for black examinees as measured by the verbal items.

The foregoing conclusions have been rather enthusiastic. The bias indices, especially the SOS statistics yield consistent results (with these sample sizes). They show appreciable discriminant validity between the biased and nonbiased studies. And, when the test questions themselves are examined, the indices seem to have signaled interpretable instances of differential performance. This enthusiasm must be tempered somewhat by the following result. In practical terms we wished to quantify the effect of having biased items in the test. Therefore, we rescored the math test deleting the seven items found to be consistently biased against blacks. We compared the new black and white means in the metric of the white standard deviation. The difference was .81σ. For the unexpurgated test it had been .91σ. The effect of the biased items (however consistent) is small but not trivial.[7] The relatively small magnitude of the bias effect can also be seen by examining the ICC graphs for typical biased items. Although the curves are discernably different, the probability differences are not very large. Item 6, comparison 2, was selected for illustration (Figure 4a) because it had the very largest unsigned area statistic of all the biased items. At its height the probability difference between blacks and whites is .13. More typically the largest black-white difference on a biased item is only .05 to .10. This would mean on average roughly one more item correct for blacks if the biased items were removed.

---

[7]The effect on black-white differences would have been smaller still, if we had deleted the three items biased against whites. However, the bias against whites was much less interpretable.

To illustrate further the practical import of the seven items biased against blacks, we also simulated the effect on failure rates if the test had been used to make pass-fail decisions as in a minimum-competency testing program. To establish comparable cut-off scores, raw scores were selected that would fail 10% of the whites on both the full and debiased tests. The corresponding failure rates for blacks on the two tests were 36.3% and 30.3% respectively.

The finding that the overall effect of bias is small tempers both our methodological and substantive conclusions. To be sure, we must

remember that internal bias indices cannot detect constant bias. Because the format of all the HSB problems requires some verbal reasoning we may have underestimated the effect of pervasive bias from this source. It is also plausible that a math achievement instrument developed for a national survey would be much less biased than many other tests. Because the bias results were consistent and interpretable in a test with a relatively small bias effect, we are inclined to believe that the indices are sensitive to relatively subtle but meaningful sources of bias. We expect that the desirable properties of the indices for bias detection would be enhanced in situations where there was a greater amount of bias. We would predict for example, that in field trials of new test items, there would be more bias to be detected.

Bias Results for the Vocabulary Test

Bias indices for the Vocabulary Test are presented in Table 8. Again, Comparisons 1 and 2 are randomly equivalent black-white analyses. Comparison 3 is between two random samples of whites, a circumstance where there should be no bias. The largest values obtained in the white-white comparison were used as baselines for interpreting the size of indices in the between-ethnic comparisons. Because two items in the white-white analysis stood out as different from the typical range of values, the indices from the second-most discrepant item were used to establish the cut-offs.

The methodological results from the Vocabulary Test were discussed earlier. Generally, they corroborated the findings based on the Math Test, but patterns were sometimes weaker because there was overall less internal bias in the Vocabulary Test. This test was very difficult for both groups.

Table 8

Signed and Unsigned Bias Indices for Vocabulary Items
(in three Comparison Studies)

	Comparison 1: W4, B4					Comparison 2: W5, B5					Comparison 3: W4, W5				
	Unsigned			Signed		Unsigned			Signed		Unsigned			Signed	
	UA	SOS2	χ^2	SA	SOS4	UA	SOS2	χ^2	SA	SOS4	UA	SOS2	χ^2	SA	SOS4
Item 1															
2	.26*	11.04*	23.31*	-.26*	-11.04*	.19	10.74*	15.00*	-.18	-10.73*	.05	.23	1.56	-.05	-.15
3	.02	.45	.43	.02	.44	.06	.74	1.70	.06	.74	.04	1.54	.96	.01	1.53
4	.29*	10.92*	17.66*	.15	9.77*	.32*	13.25*	23.22*	.32*	13.25*	.19	2.90	7.67*	-.13	-.12
5	.07	6.36	4.24	-.06	-6.35	.07	1.05	2.12	-.07	-1.05	.05	1.03	.58	.02	-1.00
6	.29*	7.06	16.09*	-.21	-6.82	.33*	12.32*	23.05*	-.33*	-12.32*	.14	2.51	4.99	.13	2.51
7	.02	.03	.06	.00	.00	.07	.64	1.61	.07	.63	.03	.25	.38	.02	.22
8	.42*	3.94	8.89*	.42*	3.94	.29*	21.89*	7.19*	.01	21.76*	.24*	7.42*	7.19*	.14	-7.00*
9	.10	1.83	2.72	-.10	1.83	.29*	4.09	9.98*	.16	-1.40	.14	2.12	4.78	-.04	.84
10	.19	2.62	5.12	-.11	-2.09	.21	3.11	6.14	-.06	.39	.03	.07	.21	.02	.07
11	.15	7.26	7.49*	-.11	-7.24*	.15	6.01	5.31	-.06	-5.98	.03	.16	.29	.03	-.13
12	.31*	13.57*	10.28*	-.08	12.23*						.25*	21.69*	11.79*	.00	21.24*
13	.28*	3.71	5.22	-.25*	-3.71										
14	.18	1.68	3.21	-.18	-1.68	.14	1.98	2.98	-.13	-1.98	.10	.15	1.22	-.09	-.15
15	.21	3.46	6.45	.21	3.46	.18	5.71	5.99	.04	5.48	.12	1.09	3.76	.10	1.05
16	.20	8.99*	14.46*	-.07	8.90*	.39*	7.58*	34.99*	-.30*	6.84	.24*	.60	5.99	.23	-.28
17	.73*	28.10*	39.76*	.73*	28.10*	.69*	26.07*	53.31*	.67*	26.03*	.90	.34	.58	-.01	.30
18	.24*	21.44*	35.29*	.02	15.90*	.17	8.08*	13.94*	-.01	4.75					
19	.23	2.07	3.95	.08	1.58	.17	2.68	3.79	.16	2.67	.05	.25	.53	-.05	-.24
20	.08	1.00	2.37	-.08	-1.00										
21						.17	5.33	.77	.17	5.33	.24*	.13	2.65	-.24*	-.08
22	.27*	1.32	4.96	.27*	1.27	.03	.04	.21	.03	.04	.10	1.12	2.10	-.04	.87
23	.03	.12	.31	-.03	-.12	.08	.51	1.39	.06	.49	.09	.40	1.09	-.03	-.07
24	.34*	1.07	3.75	.22	-.21	.20	1.21	1.88	-.12	.95	.13	3.51	3.58	.13	3.51
25	.14	.88	.90	-.06	.82	.04	.02	.08	.04	.01	.02	.13	.13	.02	.13
26	.22	6.93	10.82*	.09	-6.57										
27	.22	6.12	8.55*	.09	-5.78										

Note: To establish a baseline for judging the magnitude of the bias indices, the values from the second most deviant item in the white-white comparison were used. Indices that exceeded this cut-off in other comparisons are starred as "biased." For the sake of consistency the item 8, W1,W2,χ^2 of 7.19 was used, however 5.99 is the critical value for statistical significance at $\alpha = .05$.

Inspection of the content also suggested that the test was extremely unidimensional; for example, we could not categorize the words à priori as being more or less frequent in everyday language. All of the words seemed to have a literary flavor and were school and book oriented. Therfore, we were uncertain as to whether the analysis would detect differential difficulty.

The consistently biased items seen in Table 8 are not immediately interpretable. Initially we conjectured that there might be some speed effects present in this test since the two parts had time limits of only 5 and 4 minutes respectively. (Note that Part II starts with item 16.) However, there were not, in fact, appreciably different omitted or not-reached rates between the two groups. Four items appear to be consistently biased against blacks, items 4, 16, 17, 18. This result was puzzling because these are consistently the easiest items in the test. Only three other items (#1, 3 and 5) are as easy (and #1 could not be estimated). Apparently there may be a floor effect here whereby blacks scoring near chance on many other items in the test cannot look as different on the very difficult items as they do on easy items. (Note, item 8 would have contradicted this trend since it is biased against blacks and is very difficult (P_W = .35; P_B = .20), however, we ignored item 8 because it was also "biased" in the white-white comparison.)

Summary and Conclusions

The purpose of this research was to apply item response theory bias detection procedures to both a mathematics achievement and vocabulary test. Because the results of previous item-bias studies have often been uninterpretable, we wished to account for statistical artifacts by conducting

cross-validation or replication studies. Therefore, each analysis was repeated on randomly equivalent samples of blacks and whites. Furthermore, to establish a baseline for judging bias indices that might be attributable only to sampling fluctuations, bias analyses were conducted comparing randomly selected groups of whites. Also, to assess the effect of mean group differences on the appearance of bias, pseudo-ethnic groups were created. That is, samples of whites selected to simulate the average black-white difference were also tested for bias.

The validity and sensitivity of the IRT bias indices was supported by several findings:

1. A relatively large number of items (10 of 29) on the Math Test were found to be consistently biased, i.e., the results were replicated in parallel analyses. (Seven were biased against blacks, three were biased against whites.)
2. The bias indices were substantially smaller in white-white analyses. That is, with the exception of one or two estimation artifacts, indices did not find bias in situations of no bias.
3. Furthermore, the indices (with the possible exception of χ^2) did not find bias in the pseudo-ethnic comparison. Therefore, bias by these methods is not an artifact of mean-group differences.
4. The pattern of between study correlations showed high consistency between analyses where bias was plausibly present (e.g., between parallel ethnic comparisons).
5. Also, the indices met the discriminant validity test. That is, the correlations were low between conditions where bias should not be present.

6. For the math test where a substantial number of items appeared biased, the results were interpretable. Verbal math problems were systematically biased against blacks.
7. Finally, the desirable pattern of between comparison correlations was replicated on the Vocabulary Test, albeit somewhat weaker because of less bias on this measure.

Overall the sums-of-squares statistics (weighted by the inverse of the variance errors) were judged to be the best indices for quantifying ICC differences between groups. Not only were these statistics the most consistent in detecting bias in the ethnic comparisons, they also intercorrelated the least in situations of no bias. Lord's (1980) asymptotic chi-square was consistent but was sometimes sensitive to parameter differences that did not have corresponding effects on ICC differences.

When statistically biased items on the Math Test were examined a strong relationship was found between the verbal properties of the item and bias classification. Most of the verbal problems on the test were biased against blacks, and with one exception numeric problems were not. This highly reliable and interpretable result had to be tempered by the finding that the magnitude of the bias effect was relatively small. When items biased against blacks were deleted and the test rescored, the difference between blacks and whites was changed from .91σ to .81σ. The bias indices are apparently sensitive to consistent but subtle effects. Presumably the validity evidence for the bias statistics would be increased in situations where there is greater bias, as in field tests of newly developed test items. We did not make substantive interpretations of bias findings for the Vocabulary Test. The amount of internal bias was much less for this instrument.

References

Angoff, W. H. A technique for the investigation of cultural differences. Paper presented at the annual meeting of the American Psychological Association, Honolulu, September 1972.

Angoff, W. H. Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), Handbook of Methods for Detecting Test Bias. Baltimore, Md.: Johns Hopkins University Press, 1982.

Berk, R. A. (Ed.). Handbook of Methods for Detecting Test Bias. Baltimore, Md.: Johns Hopkins University Press, 1982.

Bond, L. Bias in mental tests. In B. F. Green (Ed.), Issues in Testing: Coaching, Disclosure and Ethnic Bias. San Francisco: Jossey Bass, 1981.

Bougon, M. G., & Lissak, R. I. Detecting item bias using the three-parameter logistic model and the B and C statistics. Paper presented at the annual meeting of the American Psychological Association, Los Angeles, August 1981.

Carroll, J. B., Davies, P., Richman, B. The American Heritage Word Frequency Book. Boston: Houghton Mifflin, Co., 1971.

Cole, N. S. Bias in testing. American Psychologist, 1981, 36, 1067-1077.

Divgi, D. R. Potential pitfalls in applications of item response theory. Paper accompanying discussant comments at the annual meeting of the National Council on Measurement in Education, Los Angeles, April 1981.

Green, D. R., Coffman, W. E., Lenke, J. M.; Raju, N. S., Handrick, F. A., Loyd, B. H., Carlton, S. T., & Marco, G. L. Methods used by test publishers to debias standardized tests. In R. A. Berk (Ed.), Handbook of Methods for Detecting Test Bias. Baltimore, Md.: Johns Hopkins University Press, 1982.

Green, D. R., & Draper, J. F. Exploratory studies of bias in achievement tests. Paper presented at the annual meeting of the American Psychological Association, Honolulu, September 1972.

Hunter, J. E. A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement test items. Paper presented at the National Institute of Education Conference on Test Bias, Maryland, December 1975.

Ironson, G. H. Use of chi-square and latent trait approaches for detecting item bias. In R. A. Berk (Ed.), Handbook of Methods for Detecting Test Bias. Baltimore, Md.: Johns Hopkins University Press, 1982.

Ironson, G. H., & Subkoviak, M. A comparison of several methods of assessing item bias. Journal of Educational Measurement, 1979, 16, 209-225.

Jensen, A. R. An examination of cultural bias in the Wonderlic Personnel Test. Intelligence, 1977, 1, 51-64.

Jensen, A. R. How biased are culture-loaded tests? Genetic Psychology Monographs, 1974, 90, 185-244.

Jensen, A. R. Test bias and construct validity. Phi Delta Kappan, 1976, 58, 340-346.

Linn, R. L. Selection bias: multiple meanings. Paper based as a presedential address to the Division of Evaluation and Measurement at the annual meeting of the American Psychological Association, August 1982.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. An investigation of item bias in a test of reading comprehension, 1980. (ERIC Document Reproduction Service No. ED 184 091).

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. Item bias in a test of reading comprehension. Applied Psychological Measurement, 1981, 5, 159-173.

Lord, F. M. A study of item bias using item characteristic curve theory. In Y. H. Poortinga (Ed.), Basic Problems in Cross-Cultural Psychology. Amsterdam: Swets and Zeitlinger, 1977.

Lord, F. M. Applications of Item Response Theory to Practical Testing Problems. Hillsdale, N. J.: Erlbaum, 1980.

Merz, W. R., & Grossen, N. E. An empirical investigation of six methods for examining test item bias. Report submitted to the National Institute of Education, Grant NIE-6-78-0067, California State University, Sacramento, Calif., 1979.

Petersen, N. S. Bias in the selection rule: Bias in the test. Paper presented at the Third International Symposium on Educational Testing, University of Leyden, The Netherlands, June 1977.

Petersen, N. S., & Novick, M. R. An evaluation of some models for culture-fair selection. Journal of Educational Measurement, 1976, 13, 3-29.

Plake, B. S. A comparison of a statistical and subjective procedure to ascertain item validity: One step in the test validation process. Educational and Psychological Measurement, 1980, 40, 397-404.

Reckase, M. D. Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 1979, 4, 207-230.

Rudner, L. M. Biased item detection techniques. Journal of Educational Statistics, 1980, 5.

Rudner, L. M., Getson, P. R., & Knight, D. L. A Monte Carlo comparison of seven biased item detection techniques. Journal of Educational Measurement, 1980, 17, 1-10.

Sandoval, J., & Miille, W. P. W. Accuracy of judgments of WISC-R item difficulty for minority groups. Journal of Consulting and Clinical Psychology, 1980, 48, 249-253.

Scheuneman, J. A method of assessing bias in test items. Journal of Educational Measurement, 1979, 16, 143-152.

Scheuneman, J. D. A posteriori analyses of biased items. In R. A. Berk (Ed.), Handbook of Methods for Detecting Test Bias. Baltimore, Md.: Johns Hopkins University Press, 1982.

Shepard, L. A. Definitions of bias. In R. A. Berk (Ed.), Handbook of Methods for Detecting Test Bias. Baltimore, Md.: Johns Hopkins University Press, 1982.

Shepard, L. A. Identifying bias in test items. In B. F. Green (Ed.), Issues in Testing: Coaching, Disclosure and Ethnic Bias. San Francisco: Jossey Bass, 1981.

Shepard, L. A., Camilli, G., & Averill, M. Comparison of procedures for detecting test-item bias with both internal and external ability criteria. Journal of Educational Statistics, 1981, 6, 317-375.

Wood, R. L., & Lord, F. M. A User's Guide to LOGIST. Research Memorandum. Princeton, N. J.: Educational Testing Service, 1976.

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST: A Computer Program for Estimating Examinee Ability and Item Characteristic Curve Parameters. Research Memorandum. Princeton, N.J.: Educational Testing Service, 1976.